ОРИГИНАЛЬНАЯ СТАТЬЯ

DOI: 10.26794/2587-5671-2023-27-1-103-115
УДК 336.7(045)
JEL G21, G29

Комбинированная схема отбора признаков для разработки банковских моделей

С.В. Афанасьев[a], Д.М. Котерева[b], А.А. Мироненков[c], А.А. Смирнова[d]
[a, b, d] Национальный исследовательский университет «Высшая школа экономики», Москва, Россия;
[a, b, d] КБ «Ренессанс Кредит», Москва, Россия;
[c] Московский государственный университет имени М.В. Ломоносова, Москва, Россия

АННОТАЦИЯ

Методы машинного обучения успешно применяются в самых разных областях банковского кредитования. За годы их применения банки накопили огромные массивы данных о заемщиках. С одной стороны, это позволило более точно предсказывать поведение заемщика, с другой — породило проблему избыточности данных, которая сильно усложняет разработку моделей. Чтобы решить эту проблему применяют методы отбора признаков, позволяющие повысить качество моделей. Эти методы делятся на три типа: фильтры, обертки и вложения. Фильтры являются простыми и быстрыми методами, при использовании которых можно находить одномерные зависимости. Обертки и вложения позволяют более качественно проводить отбор признаков, поскольку учитывают многомерную зависимость, однако данные методы требуют значительных вычислительных ресурсов и могут плохо работать на больших высокоразмерных выборках. В данной статье авторы предлагают схему комбинированного отбора признаков CFSS, в которой на первых этапах отбора используются грубые фильтры, а на финальных — обертки для более качественного отбора. Такая архитектура повышает качество и скорость отбора признаков на больших высокоразмерных выборках для промышленного моделирования в банковских задачах. Проведенные авторами эксперименты для четырех типов банковских задач (анкетный скоринг, поведенческий скоринг, отклик клиентов на кросс-сейл предложение и взыскание просроченной задолженности) показали, что предложенный метод работает лучше, чем классические методы, содержащие только фильтры или только обертки.

Ключевые слова: отбор признаков; машинное обучение; значимость переменных; фильтры; обертки; вложения

Для цитирования: **Афанасьев С.В., Котерева Д.М., Мироненков А.А., Смирнова А.А. Комбинированная схема отбора признаков для разработки банковских моделей. *Финансы: теория и практика.* 2023;27(1):103-115. DOI: 10.26794/2587-5671-2023-27-1-103-115**

ORIGINAL PAPER

Combined Feature Selection Scheme for Banking Modeling

S.V. Afanasyev[a], D.M. Kotereva[b], A.A. Mironenkov[c], A.A. Smirnova[d]
[a, b, d] National Research University Higher School of Economics, Moscow, Russia;
[a, b, d] Renaissance Credit Bank, Moscow, Russia;
[c] Lomonosov Moscow State University, Moscow, Russia

ABSTRACT

Machine learning methods have been successful in various aspects of bank lending. Banks have accumulated huge amounts of data about borrowers over the years of application. On the one hand, this made it possible to predict borrower behavior more accurately, on the other, it gave rise to the problem a problem of data redundancy, which greatly complicates the model development. Methods of feature selection, which allows to improve the quality of models, are apply to solve this problem. Feature selection methods can be divided into three main types: filters, wrappers, and embedded methods. Filters are simple and time-efficient methods that may help discover one-dimensional relations. Wrappers and embedded methods are more effective in feature selection, because they account for multi-dimensional relationships, but these methods are resource-consuming and may fail to process large samples with many features. In this article, the authors propose a combined feature selection scheme (CFSS), in which the first stages of selection use coarse filters, and on the final — wrappers for high-quality selection. This architecture lets us increase the quality of selection and reduce the time necessary to process large multi-dimensional samples, which are used in the development

of industrial models. Experiments conducted by authors for four types of bank modelling tasks (survey scoring, behavioral scoring, customer response to cross-selling, and delayed debt collection) have shown that the proposed method better than classical methods containing only filters or only wrappers.
Keywords: feature selection; machine learning; feature importance; filters; wrappers; embedded methods

For citation: **Afanasyev S.V., Kotereva D.M., Mironenkov A.A., Smirnova A.A. Combined feature selection scheme for banking modeling. *Finance: Theory and Practice*. 2023;27(1):103-115. DOI: 10.26794/2587-5671-2023-27-1-103-115**

ВВЕДЕНИЕ

Методы машинного обучения успешно применяются в самых разных областях банковского кредитования. Огромные массивы данных позволяют более точно предсказывать поведение заемщика, при этом порождают проблему избыточности данных, что усложняет разработку моделей и может приводить к неудовлетворительным результатам. Чтобы решить эту проблему, были предложены различные методы отбора признаков [1]. Основная концепция этих методов состоит в том, чтобы уменьшить размерность признакового пространства за счет исключения избыточных признаков.

Предлагаемые в научной литературе методы отбора признаков делятся на три типа: фильтры (Filter methods), обертки (Wrapper methods) и вложения (Embedded methods) [2].

Большинство предлагаемых в научной литературе методов тестируются на открытых репозиториях, которые содержат либо мало наблюдений (от нескольких десятков до нескольких тысяч), либо мало признаков (несколько десятков) [3–5]. При этом на практике банковского моделирования используются выборки, на порядки больше научных баз данных, и содержат от нескольких сотен тысяч до нескольких миллионов наблюдений и от нескольких сотен до нескольких тысяч признаков. На таких выборках предлагаемые в исследованиях методы либо не дают декларируемого результата, либо работают очень долго. Для решения этих проблем мы предлагаем метод комбинированного отбора признаков (Combined Feature Selection Scheme — CFSS), который представляет из себя гибридную многоступенчатую схему отбора, где на первых этапах применяются фильтры, а на последующих — обертки. В качестве фильтров мы используем методы очистки данных, проверку стабильности признаков, корреляцию признаков с целевой переменной, матрицу взаимных корреляций [6] и VIF анализ [7], в качестве оберток — пермутационный метод на основе случайного леса (Random Forest) [8] и оценку p-value признаков с помощью метода обратного исключения (Backward Elimination) [9].

Предложенный нами метод был протестирован на четырех выборках для различных банковских задач: предсказание вероятности просрочки по кредиту в момент оформления кредитной заявки (Application PD), предсказание вероятности будущей просрочки по кредиту в процессе жизни кредита (Behavioral PD), оценка вероятности отклика клиента на рекламное предложение (CRM PTB) и оценка вероятности перехода просрочки по кредиту в более поздний месяц (Collection Allocation).

Полученные результаты показали, что метод CFSS хорошо работает на больших высокоразмерных выборках. Мы также продемонстрировали, что за счет гибкого комбинирования фильтров и оберток метод CFSS позволяет добиваться более высокой обобщающей способности моделей, чем методы обертки без использования фильтров. Дополнительно проведенные эксперименты показали, что метод CFSS работает в десятки раз быстрее классических методов отбора признаков, что является важным преимуществом метода в промышленном применении.

ОБЗОР ЛИТЕРАТУРЫ

Фильтры

К наиболее простым методам отбора признаков относятся фильтры, которые позволяют отбирать признаки независимо от разрабатываемой модели.

Отбор переменных с помощью матрицы корреляции (CFS) позволяет оценивать подмножества признаков, исходя из гипотезы, что хорошие поднаборы содержат такие признаки, которые не коррелируют друг с другом, но при этом сильно коррелируют с целевой переменной. Простейший способ выделить сильно коррелированные признаки заключается в построении матрицы парной корреляции признаков. Данный подход получил широкое распространение на практике. К плюсам метода можно отнести простоту реализации и интерпретации. Среди недостатков метода можно выделить чувствительность к качеству данных (выбросам, ошибкам и т.п.), а также неспособность выявлять многофакторные взаимосвязи.

Метод главных компонент (PCA), предложенный Карлом Пирсоном в 1901 г. [10] и до сих пор являющийся популярным методом в прикладных задачах, позволяет уменьшать размерность путем вычисления главных компонент матрицы признаков и последующем понижении размерности матрицы через ее сингулярное разложение [11]. Среди плюсов метода

можно отметить простоту его реализации. Среди недостатков метода PCA можно выделить чувствительность к масштабу, сложности с выбором порога отсечения для главных компонент, а также то, что в PCA не учитывается целевая переменная, из-за чего главные компоненты могут оказаться не информативными.

Несмотря на указанные недостатки, фильтры активно используются на практике и до сих пор являются предметом научных исследований. Чжан и соавторы [12] используют t-тест Уэлча для разработки алгоритмов ранней компьютерной диагностики болезни Альцгеймера.

Роффо и Мельци [5] предлагаю метод для отбора признаков на основе анализа графа, где в качестве вершин графа выступают исследуемые признаки, а в качестве ребер — сила связи между признаками. Авторы предполагают, что собственный вектор при максимальной главной компоненте в матрице смежности графа будет содержать отранжированные по важности признаки. Если в качестве функции связи взять коэффициенты линейной корреляции между признаками, то матрица смежности графа превращается в стандартную матрицу корреляций.

Обертки

Среди методов оберток наибольшую популярность получили методы ступенчатой регрессии: метод прямого отбора (Forward Selection) [13], метод обратного исключения (Backward Elimination) [13] и метод последовательного отбора (Stepwise)[1] [14]. Несмотря на свою простоту и эффективность методы ступенчатой регрессии подвергались критике в научном сообществе [15].

В научной литературе также большое внимание уделяется метаэвристическим алгоритмам оптимизации для отбора признаков, к которым относятся: алгоритм оптимизации роя частиц (PSO) [16], оптимизация на основе поведения серого волка (GWO) [14], генетический алгоритм (GA) [17, 18], оптимизация пчелиного роя (BSO) [19] и др. Шен и Чжан предложили улучшенный двухэтапный алгоритм оптимизации серого волка (Improved Gray Wolf Optimization — IGWO) [3], где на первом этапе авторы предлагают использовать вложенный метод регуляризации LASSO, а на втором — метод оптимизации поведения серого волка (GWO). Басак и соавторы предложили метод обертку Reinforced Swarm Optimization (RSO) [4], который представляет из себя улучшенный алгоритм оптимизации пчелиного роя, где вместо BSO-оптимизации используется подход на основе обучения с подкреплением. Методы отбора признаков на основе метаэвристических алгоритмов широко используются для выбора хорошего приближения в различных сложных задачах оптимизации, однако они не всегда обеспечивают выбор наилучшего решения в силу того, что обучение итоговых моделей может проводиться с помощью других алгоритмов машинного обучения, для которых выбранные признаки могут оказаться не оптимальными.

Среди других эффективных оберток выделяют пермутационные методы на основе случайного леса [18, 20]. В пермутационных методах оцениваемые признаки не удаляются из выборки, т.е. признаковое пространство остается несмещенным. К важным преимуществам пермутационных методов, основанных на алгоритмах случайного леса, относят возможность получения несмещенных оценок важностей за счет использования рандомизированных деревьев (в отличии от градиентного бустинга, где деревья зависимы и оценки получаются смещенными). Среди недостатков данных методов выделяют высокую вычислительную сложность, что накладывает ограничения на применимость данных методов для больших высокоразмерных выборок. Силке и соавторы [8] попытались решить эту проблему, предложив пермутационный метод New Approach, который показал высокую эффективность по скорости работы на больших выборках. Однако авторы также продемонстрировали, что пермутационные методы плохо работают на высокоразмерных выборках, когда количество признаков исчисляется несколькими тысячами и более.

Вложения

Регуляризация относится к группе вложенных методов, где отбор признаков становится частью процесса построения модели. В логистической регрессии, ставшей банковским стандартом [21], наиболее распространенными методами регуляризации являются L1 (LASSO) [22] и L2 (Ridge) [23] (регуляризация Тихонова [24]). Общая концепция регуляризации заключается в добавлении штрафного слагаемого к функционалу ошибки, которое наказывает модель за чрезмерную сложность. Регуляризация L1 позволяет обнулять часть весовых коэффициентов регрессии, а регуляризация L2 ограничивает их норму [25]. Регуляризация L1 имеет ряд недостатков и плохо работает на данных большой размерности с небольшим количеством наблюдений. Хуэй и Тревор [26] предлагают обойти

[1] SAS Institute Inc. (1989) SAS/STAT User's Guide, Version 6, Fourth Edition, Vol. 2, Cary, NC. URL: https://www.scirp.org/reference/ReferencesPapers.aspx?ReferenceID=1542754 (дата обращения: 07.02.2023).

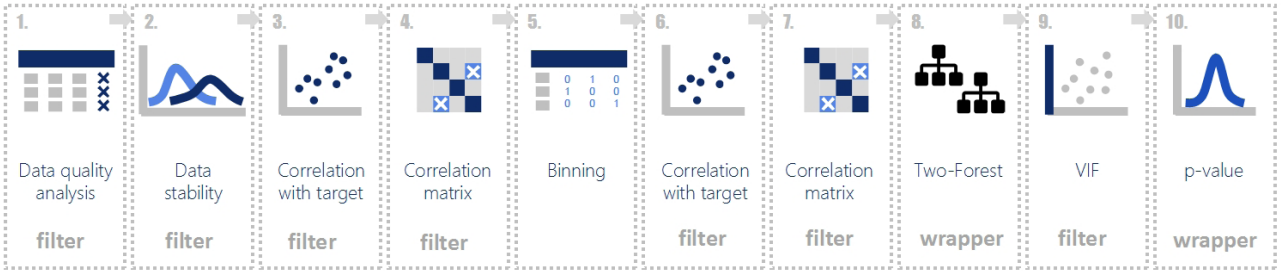

Рис. 1 / Fig. 1. **Схема комбинированного отбора признаков / Combined feature selection scheme**

Источник / Source: составлено авторами / compiled by the authors.

эти ограничения с помощью подхода Elastic Net — комбинации методов L1 и L2.

СХЕМА КОМБИНИРОВАННОГО ОТБОРА ПРИЗНАКОВ

Описанные в научной литературе плюсы и минусы накладывают ограничения на применимость предлагаемых методов отбора признаков к практическим бизнес-задачам. Если нужны быстрые методы с низкими требованиями к вычислительным ресурсам, то фильтры будут наиболее оптимальными. При этом для более высокого качества моделей необходимо использовать методы обертки и вложения, для которых могут потребоваться большие вычислительные мощности. Важно отметить, что полученные в научных исследованиях результаты могут не воспроизводиться на практике, на больших высокоразмерных данных. Эти проблемы мотивировали нас разработать метод комбинированного отбора признаков, который включает в себя 10 шагов работы с данными (*рис. 1*).

Анализ качества данных

Качество данных — обобщенное понятие, отражающее степень пригодности данных к решению определенной задачи. Среди методов анализа качества данных можно выделить:

1) разведочный анализ (Exploratory Data Analysis, EDA) [27] — выявление основных свойств данных, поиск общих закономерностей, анализ распределений, выбросов и т.д.;

2) анализ пропусков и неполноты данных — проводится с использованием статистических показателей, таких как количество непустых наблюдений, пропусков, минимальное и максимальное значения, медиана, модальное значение, стандартное отклонение, квантили и др.;

3) анализ аномалий — статистический и экспертный анализ причин появления наблюдений, выходящих за пределы допустимого диапазона значений переменной. Основные методы работы с аномалиями сводятся к построению распределения по наблюдаемой переменной и последующему определению пороговых значений в «хвостах» распределения. Также используются альтернативные методы работы с аномалиями, такие как монотонная трансформация переменных (логарифмическая и др.), расчет z-score и др.[2]

Анализ стабильности и непрерывности данных

Для статистических алгоритмов необходимо, чтобы данные были непрерывными и стабильными. Причиной нестабильности в данных может быть изменение бизнес-процессов банка, законодательства, клиентского поведения, форматов данных и т.д.

Перед оценкой стабильности все строковые переменные необходимо преобразовать в числовой формат с помощью метода LabelEncoder[3] (пропуски заменяются уникальным числовым значением). Для оценки стабильности признаков необходимо рассчитать дивергенции между распределениями, построенными на разных временных периодах. Для этого проводится оценка стабильности на разных периодах:

1) *большие периоды:* выборка делится на равные подвыборки с большим интервалом (например, по полугодиям), после чего на этих подвыборках попарно сравниваются распределения признаков по принципу «каждый с каждым» (*рис. 2а*);

2) *маленькие периоды:* общая выборка разбивается на равные маленькие подвыборки (например, по месяцам), после чего попарно сравниваются распределения признаков по смежным периодам (*рис. 2б*).

В схеме CFSS для расчета дивергенции между подвыборками рассчитываются и усредняются значения трех метрик: S-статистики [28], индекса стабильности популяции (PSI) [29] и статистики Колмогорова-Смирнова (KS) [30].

---

[2] Understanding Statistics. Graham J. G. Upton, Ian T. Cook. Oxford University Press, 1996. URL: https://books.google.ru/books?id=vXzWG09_SzAC&printsec=frontcover&hl=ru#v=onepage&q&f=false (дата обращения: 30.01.2023).

[3] URL: scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html (дата обращения: 30.01.2023).

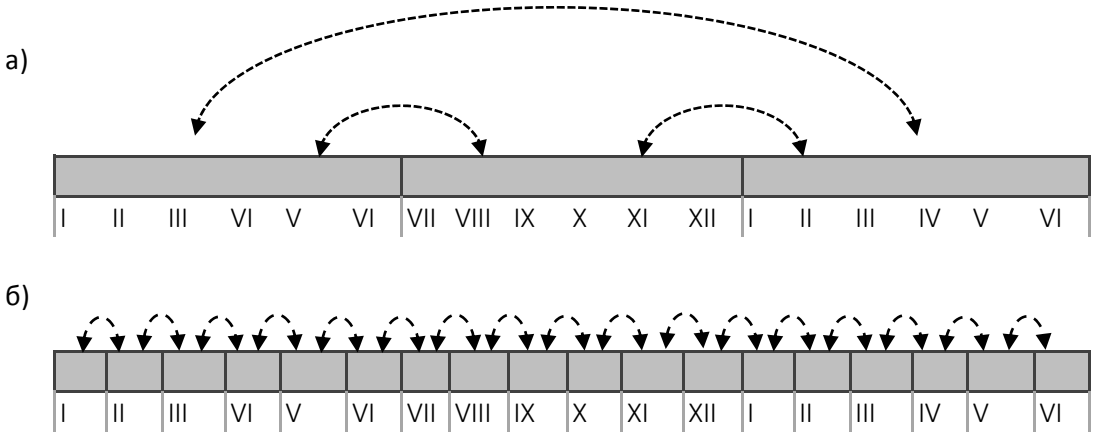

Рис. 2 / Fig. 2. **Пример разбиения выборки на периоды для оценки стабильности переменной а) большие периоды; б) маленькие периоды / An Example of a Split into Periods to Estimate the Stability of Feature a) large Periods; b) Small Periods**

Источник / Source: составлено авторами / Compiled by the authors.

Примечание / Note: месяцы обозначены римскими цифрами / Months are indicated by Roman numerals.

Данная методика позволяет выявлять как долгосрочные изменения признаков (нестабильность по большим периодам), так и частые краткосрочные изменения (нестабильность по маленьким периодам).

Корреляция признаков с целевой переменной

Анализ корреляции признаков с целевой переменной позволяет отобрать признаки, которые сильно влияют на целевую переменную. При этом данный метод не учитывает сложные зависимости между признаками, поэтому его можно отнести к «грубым» методам фильтрам, которые можно применять для первичного отбора признаков.

Методы анализа корреляции зависят от типа целевой переменной и типа исследуемого признака[4].

Признаки, которые не проходят тест на значимость корреляции с целевой переменной, исключаются из дальнейшего анализа. Пороги значимости задаются как эвристики или выбираются экспериментально.

В комбинированной схеме CFSS проверка признаков на корреляцию с целевой переменной выполняется два раза — до бинаризации признаков и после (шаг 6, *рис. 1*).

Матрица корреляций

Сильно коррелированные признаки можно выявлять с помощью матрицы корреляции (CFS), которая имеет вид:

$$R_x = \begin{pmatrix} 1 & r_{x_1x_2} & \dots & r_{x_1x_n} \\ r_{x_2x_1} & 1 & \dots & r_{x_2x_n} \\ \dots & \dots & \dots & \dots \\ r_{x_nx_1} & r_{x_nx_2} & \dots & 1 \end{pmatrix}, \quad (1)$$

где $r_{x_ix_j}$ — корреляция между i-м и j-м признаками.

Для непрерывных признаков рассчитывается корреляция Пирсона, а для категориальных и бинарных признаков — корреляция Спирмена.

В комбинированной схеме CFSS отбор признаков с помощью матрицы корреляций выполняется два раза — до бинаризации признаков и после (шаг 7, *рис. 1*). Пороговые значения для отбора признаков задаются как эвристики и зависят от типа задачи. Для наших экспериментов мы выставляли следующие пороги: 90% для этапа «до бинаризации признаков» (слабая фильтрация) и 70% для этапа «после бинаризации признаков» (сильная фильтрация).

Dummy-кодирование категориальных переменных

После отбора признаков первичными фильтрами необходимо преобразовать категориальные признаки в бинарные переменные для возможности их использования в регрессионных алгоритмах. Для логистической регрессии используется процедура dummy-кодирования [31] по методу полного ранга, когда одна из категорий удаляется. Таким образом после dummy-кодирования категориальной переменной получается $k - 1$ бинарная переменная, где k — количество категорий в исходном признаке.

[4] Айвазян С.А., Мхитарян В.С. Прикладная статистика. Основы эконометрики. Учебник для вузов. В 2 т. 2-е изд., испр. Т. 1. Теория вероятностей и прикладная статистика. М.: ЮНИТИ-ДАНА; 2001. 656 с.

После трансформации категориального признака для каждой новой бинарной переменной рассчитается количество наблюдений «положительного» класса. Все бинарные переменные, для которых количество наблюдений «положительного» класса меньше заданного порога значимости, объединяются в одну категорию. Для остальных типов целевых переменных считается количество наблюдений в категории. Порог значимости задается как эвристика или подбирается экспериментально. Для наших экспериментов мы задали порог, равный 10.

Двухлесовый метод (Two-Forest)

После первичной фильтрации признаков в CFSS используются методы обертки, учитывающие сложные взаимосвязи между признаками. Методы отбора с помощью случайного леса являются наиболее точными обертками [18, 20]. Для отбора методом случайного леса необходимо оценить важность (importance) каждого признака с помощью одного из двух подходов:

1. Важность на основе уменьшения неоднородности:

1) для каждого дерева случайного леса вычисляется сумма уменьшений неоднородности на всех ветвлениях, связанных с данной переменной;

2) итоговая сумма уменьшений неоднородности делится на общее количество деревьев;

3) шаги (1) и (2) повторяются для всех переменных.

Искомая важность признака — это частота использования переменной в качестве предиктора ветвления.

2. Важность на основе уменьшения качества прогнозирования при случайной перестановке (пермутации):

1) обучается модель случайного леса;

2) на тестовом/OOB множестве рассчитывается ошибка[5];

3) фиксируется переменная (или группа переменных) и случайно переставляются ее значения на тестовом/OOB множестве;

4) вычисляется разность между ошибкой на исходном множестве и ошибкой на множестве с перестановкой.

Вычисленная разность ошибок является пермутированной важностью переменной.

В схеме CFSS используется адаптированный метод New Approach [8], который мы назвали Two-forest[6]. Общая концепция двухлесового метода заключается в оценке важности признаков как качества прогнозирования при случайной перестановке (пермутации):

$$VI_j = P\left(Y \neq f\left(X_1,\ldots,X_j^*,\ldots,X_p\right)\right) - P\left(Y \neq f\left(X_1,\ldots,X_j,\ldots,X_p\right)\right). \quad (2)$$

Адаптированный двухлесовый метод работает по следующему алгоритму:

1. Обучающая выборка репрезентативно разбивается на две равные части.

2. На каждой подвыборке обучается случайный лес[7].

3. Оценивается качество на подвыборке, на которой модель не обучалась.

4. Каждая переменная случайным образом перемешивается и считается результат для каждой из двух моделей на отложенных подвыборках.

5. Рассчитывается разница между полученном на 3-м шаге baseline-значением и новым значением.

6. Важность переменной рассчитывается как среднее значение важностей на двух подвыборках.

7. Для полученной важности рассчитывается значение p-value:

1) выбираются наблюдения с отрицательными значениями важности;

2) выбираются наблюдения с нулевыми значениями важности;

3) отрицательные значения важности умножаются на (–1);

4) векторы, полученные на шагах (1)–(3), конкатенируются;

5) для полученного вектора строится кумулятивное распределение;

6) на полученном распределении рассчитывается p-value.

8. Выбираются переменные, у которых значение p-value ниже заданного порога. Возможны следующие эвристики:

1) разделить важность на среднее значение baseline, выбираются те переменные, у которых изменение больше заданного порога;

[5] OOB (Out-of-Bag) — оценка качества для каждого наблюдения только по тем деревьям ансамбля, которые на данном наблюдении не обучались (т.е. использование тех объектов, которые не входили в состав обучающей выборки для каждого базового дерева).

[6] Мы назвали метод «Two-forests» (двухлесовый метод), поскольку в данном подходе обучается сразу два случайных леса.

[7] В зависимости от типа целевой переменной применяются разные алгоритмы: Random Forest Classifier — для бинарной целевой переменной; Random Forest Regressor — для непрерывной целевой переменной.

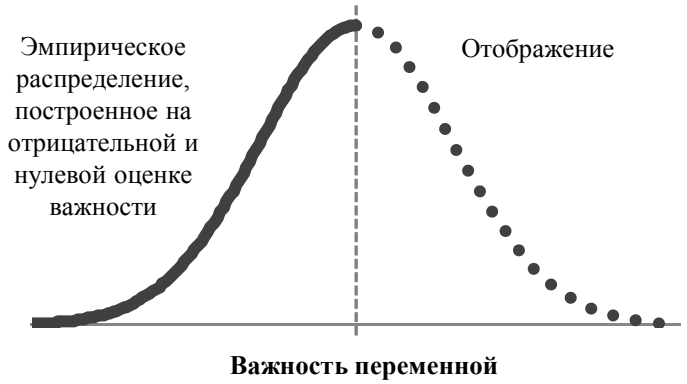

Рис. 3 / Fig. 3. **Пример построения распределения *F* на основе нерелевантных переменных (т.е. с отрицательными или нулевыми оценками важности) / Example of Constructing an F Distribution Based on Irrelevant Features (i.e. with Negative or Zero Importance Evaluations)**

Источник / Source: составлено авторами / Compiled by the authors.

2) отсортировать переменные по значениям важности и выбрать первые *N* переменных (число *N* подбирается экспериментально), значение p-value по выбранным переменным должно быть меньше 10%.

VIF анализ

Другой подход для снижения мультиколлинеарности между признаками строится на оценке показателя VIF (Variance Inflation Factor) [7]. Для расчета данного показателя необходимо построить линейную регрессию для каждого объясняющего признака (выступает в качестве целевой переменной) от всех остальных признаков. Отбор с помощью VIF анализа проводится по следующему алгоритму:

1. Для каждого признака X_i обучается линейная регрессия, в которой X_i является функцией от всех остальных признаков:

$$X_i = \beta_0 + \sum_{j=1}^{k} \beta_j X_j, \quad i \neq j, \tag{3}$$

где β_0 — свободный член регрессии;

k — общее количество признаков (включая анализируемый).

2. Рассчитывается коэффициент VIF для признака X_i:

$$VIF_i = \frac{1}{1 - R_i^2}, \tag{4}$$

где R_i^2 — коэффициент детерминации регрессии, построенной на шаге 1.

3. Проводится оценка полученных значений VIF, где применяется общее эмпирическое правило: признаки со значением VIF > 10 относятся к мультиколлинеарным [32]. Из списка мультиколлинеарных признаков удаляется признак с максимальным значением VIF.

4. Шаги 1–3 итерационно повторяются до тех пор, пока максимальное значение VIF по оставшимся признакам не станет меньше или равно 10.

Статистическая значимость признаков

Последним шагом в схеме CFSS является проверка статистической значимости признаков на основе теста отношения правдоподобия.

Процедура оценки статистической значимости признака с помощью теста отношения правдоподобия сводится к проверке нулевой гипотезы значимости признака через оценку статистики отношения правдоподобия. Для модели с вектором параметров β необходимо проверить по выборочным данным гипотезу $H_0 : g(\beta) = 0$, где $g(\beta)$ — совокупность (вектор) некоторых функций параметров. Для проверки нулевой гипотезы сравниваются функции правдоподобия полной модели (т.е. обученной на всех *n* признаках) и укороченной модели без тестируемого признака (обученной на *n* – 1 оставшихся признаках). Для этого рассчитывается статистика отношения правдоподобия (likelihood ratio test):

$$LR = 2 \cdot (L_l - L_s) = 2 \cdot \ln \frac{L_l}{L_s}, \tag{5}$$

где L_l — значение логарифмической функции правдоподобия полной модели;

L_s — значение логарифмической функции правдоподобия укороченной модели.

Статистика LR при нулевой гипотезе имеет распределения хи-квадрат с *q* степенями свободы — $\chi^2(q)$, где *q* — количество ограничений (количество исключенных признаков). Если значение данной статистики больше критического значения распределения

Таблица 1 / Table 1

Характеристики выборок для задач банковского моделирования / Characteristics of Samples for Banking Modeling Tasks

Модель / DataSet	Период обучающей и тестовой выборок / Period of train and test samples	Период отложенной выборки (OOT) / Out-of-time period (OOT)	Наблюдений, кол-во / Observations, amount	Признаков, кол-во / Features, amount	Доля миноритарного класса (bad-rate), % / Minority class percentage, %
PTB (CRM)	01.11.2019–30.01.2020	01.02.2020–28.02.2020	545 963	1222	1,28
Behavioral PD (Scoring)			1 195 466	1087	13,88
Application PD (Scoring)			793 080	423	3,60
Allocation (Collection)	01.06.2018–30.04.2019	01.05.2019–30.06.2019	256 220	162	37,19

Источник / Source: составлено авторами / Compiled by the authors.

при заданном уровне значимости, то исключенный признак считается значимым и предпочтение отдается полной модели. В противном случае исключенная переменная признается незначимой.

Пороговое значение уровня значимости p-value задается как эвристика или подбирается экспериментально. В тестируемой схеме CFSS уровень значимости p-value был установлен 0,05.

ПОСТАНОВКА ЭКСПЕРИМЕНТА

Данные

Для проведения экспериментов использовались данные крупного российского банка. Сравнение методов отбора проводилось на четырех датасетах для банковских задач бинарной классификации:

1. *CRM: PTB (probability to bay)* — модель оценки вероятности отклика клиента на кросс-сейл предложение по кредиту.
2. *Scoring: Application PD (CASH)* — модель оценки вероятности дефолта на этапе выдачи кредита (кредиты наличными для физических лиц).
3. *Scoring: Behavioral PD (CASH)* — модель оценки вероятности дефолта в течение жизни кредита с использованием поведенческой информации о предыдущих кредитных платежах клиента. Данная модель позволяет оценивать уровень кредитного риска по портфелю ссуд для формирования резервов и капитала в соответствии с требованиями международной финансовой отчетности (МСФО 9) и на основе внутренних рейтингов (IRB, Basel II).
4. Collection: Allocation — модель оценки вероятности переката просрочки по кредиту в более поздний месяц графика платежей.

Характеристики выборок представлены в *табл. 1*.

Эксперимент

Второй эксперимент проводился для сравнения популярных в банковской практике классических методов отбора и схемы CFSS с использованием двухлесового метода. Сравнение проводилось между тремя схемами отбора:

1. *Gini Scheme* — в этой схеме на шаге 8 (*рис. 1*) вместо двухлесового метода применялся отбор признаков с использованием оценок Gini (Gini > 5%) для однофакторных логистических регрессий (все остальные этапы отбора общей схемы остались без изменений).
2. *Forward Scheme* — в этой схеме на шаге 8 применялся отбор признаков с использованием метода прямого отбора (Forward Selection).
3. *CFSS* — комбинированный отбор с использованием двухлесового метода (*рис. 1*).

Перечисленные методы оценивались в составе 10-этапной схемы комбинированного отбора, чтобы не сравнивать заведомо слабые методы Gini и Forward с сильным двухлесовым методом.

В рамках второго эксперимента также оценивалось время работы методов Forward и Two-Forest.

РЕЗУЛЬТАТЫ

Цель эксперимента заключалась в сравнении схемы CFSS с методами отраслевого стандарта. Сравнива-

Таблица 2 / Table 2

Сравнение трех схем отбора признаков: Gini, Forward, CFSS / Comparison of Three Feature Selection Schemes: Gini, Forward, CFSS

Scheme	Models	Features, amount	Gini			
			LogReg		LightGBM	
			Test	OOT	Test	OOT
Gini Scheme (GS)	PTB (CRM)	1222	0,4157	0,4312	0,4380	0,4480
	Behavioral PD (Scoring)	1087	0,6843	0,6400	0,6904	0,6493
	Application PD (Scoring)	423	0,4051	0,3980	0,4251	0,4179
	Allocation (Collection)	162	0,6048	0,6075	0,6499	0,6494
Forward Scheme (FS)	PTB (CRM)	1222	0,4259	**0,4369**	0,4302	0,4481
	Behavioral PD (Scoring)	1087	**0,6907**	**0,6466**	**0,7068**	**0,6705**
	Application PD (Scoring)	423	**0,4164**	**0,4067**	0,4356	0,4203
	Allocation (Collection)	162	**0,6143**	0,6041	0,6418	0,6436
CFSS	PTB (CRM)	1222	**0,4332**	0,4340	**0,4401**	**0,4527**
	Behavioral PD (Scoring)	1087	0,6881	0,6439	0,7050	0,6682
	Application PD (Scoring)	423	0,4093	0,4051	**0,4390**	**0,4290**
	Allocation (Collection)	162	0,6111	**0,6085**	**0,6500**	**0,6507**
			Δ Gini			
Difference: (CFSS – GS)	PTB (CRM)	1222	**0,0174**	**0,0028**	**0,0021**	**0,0047**
	Behavioral PD (Scoring)	1087	**0,0038**	**0,0040**	**0,0146**	**0,0189**
	Application PD (Scoring)	423	**0,0042**	**0,0071**	**0,0139**	**0,0111**
	Allocation (Collection)	162	**0,0062**	**0,0011**	**0,0001**	**0,0013**
Difference: (CFSS – FS)	PTB (CRM)	1222	**0,0073**	-0,0030	**0,0100**	**0,0046**
	Behavioral PD (Scoring)	1087	–0,0025	–0,0027	–0,0018	–0,0023
	Application PD (Scoring)	423	–0,0072	–0,0016	**0,0034**	**0,0087**
	Allocation (Collection)	162	–0,0032	**0,0044**	**0,0082**	**0,0071**

Источник / Source: составлено авторами / Compiled by the authors.

лись схемы комбинированного отбора (*рис. 1*) с добавлением на шаге 8 трех разных методов отбора: оценка Gini (банковский стандарт), Forward (банковский стандарт) и Two-Forest (CFSS — наш подход).

Результаты экспериментов (*табл. 2*) показали, что схема Gini проиграла по качеству схемам Forward и CFSS. С другой стороны, схема Forward показала лучшие результаты для логистической регрессии, а схема CFSS — для градиентного бустинга в реализации LightGBM. Этот результат подтверждает тезис о том, что для выбора оберток необходимо учитывать тип алгоритма для финальной модели.

Сравнение средней разницы качества исследуемых моделей показало, что на логистической регрессии схема CFSS незначительно проиграла схеме FS (*рис. 4*), что демонстрирует хорошую устойчивость схемы CFSS к типу алгоритма для обучения финальной модели. Это может быть связано с тем, что в схеме CFSS после нелинейного метода Two-Forest используется линейный Backward метод (оценка p-value), который балансирует отбор в сторону линейных признаков.

Сравнение времени работы методов Forward и Two-Forest показало, что Two-Forest работает в десятки раз быстрее метода Forward (*табл. 3*). В данном

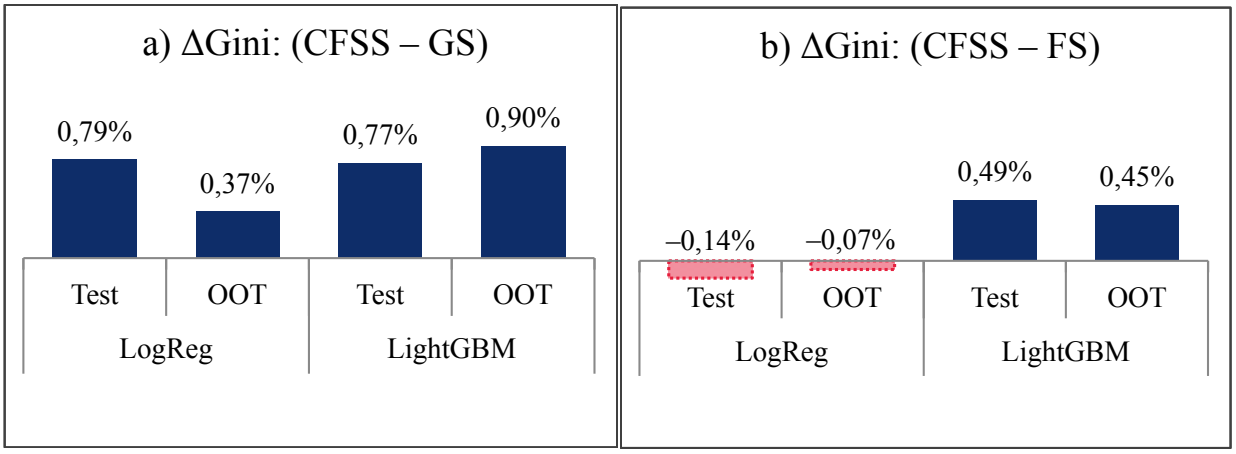

Рис. 4 / Fig. 4. **Сравнение схем Gini Scheme (GS), Forward Scheme (FS) и CFSS: a) разница Gini по моделям CFSS и GS; b) разница Gini по моделям CFSS и FS / Comparison of Gini Scheme (GS), Forward Scheme (FS) and CFSS: a) Gini Difference between CFSS and GS Models; b) Gini Difference between CFSS and FS Models**

Источник / Source: составлено авторами / Compiled by the authors.

Таблица 3 / Table 3

Время работы методов отбора Forward и Two-Forest / Selection Time of the Forward and Two-Forest Methods

Model	Observations, amount (Train)	Features, amount	Time (hh: mm: ss)		x-Times: Forward / 2Forest
			Forward	Two-Forest	
PTB (CRM)	303 220	1222	14:01:28	0:30:48	74x
Behavioral PD (Scoring)	588 385	1087	16:11:04	0:16:36	58x
Application PD (Scoring)	497 063	423	5:34:12	0:15:00	22x
Allocation (Collection)	172 250	162	3:06:40	0:05:50	32x

Источник / Source: составлено авторами / Compiled by the authors.

эксперименте сравнивалось время работы методов только на 8-м шаге общей схемы, так как все остальные этапы отбора были одинаковыми.

ВЫВОДЫ

В этой статье мы предложили схему комбинированного отбора признаков CFSS, в которой на первых этапах проводится очистка признаков и проверка их стабильности, на последующих шагах используются корреляционные фильтры, позволяющие отсеять сильно коррелированные между собой признаки, и на финальных этапах применяются методы обертки, являющиеся тонкой настройкой схемы и завершающие отбор. Такая схема отбора «от простого к сложному» позволяет сбалансировать отбор и добиться хороших результатов по качеству и скорости на больших высокоразмерных выборках.

Результаты наших экспериментов показали, что схема CFSS хорошо работает для разных типов моделей (линейных и нелинейных) и разных банковских задач (кредитный скоринг, рекламные кампании, взыскание просроченной задолженности и др.) и превосходит по качеству схемы, содержащие только фильтры или обертки.

Включение в комбинированную схему отбора нескольких методов-оберток позволяет контролировать корректность работы каждого метода на предыдущих шагах отбора. По сравнению с регрессионными подходами двухлесовый метод отбора показывает лучшее качество для нелинейных моделей и сопоставимое качество для линейных. При этом по скорости работы двухлесовый метод в десятки раз выигрывает у регрессионных методов.

Комбинированную схему отбора признаков можно полностью автоматизировать, встроив ее

в общий pipeline разработки моделей в банке. Это позволяет вести разработку моделей в режиме «End-to-End», что ускоряет процесс разработки и снижает модельные риски.

Стоит отметить, что в схеме CFSS использовался ряд фиксированных пороговых метрик, определяемых экспертом. Таким образом, схема CFSS все еще является метаэвристической, когда на некоторых этапах отбора не учитывается специфика данных. Данные эвристики, а также порядок работы методов в CFSS можно дополнительно настраивать как гиперпараметры модели, что позволит учесть специфику задачи и повысить качество финальных моделей. Однако настройка гиперпараметров приведет к увеличению временной сложности схемы CFSS. Наши будущие исследования будут посвящены проработке этих вопросов.

СПИСОК ЛИТЕРАТУРЫ / REFERENCES

1. Guyon I., Elisseeff A. An introduction to variable and feature selection. *Journal of Machine Learning Research*. 2003;3(7–8):1157–1182. DOI: 10.1162/153244303322753616
2. Hamon J. Optimisation combinatoire pour la sélection de variables en régression en grande dimension: Application en génétique animale. Docteur en Informatique Thèse. Lille: Université des Sciences et Technologie de Lille; 2013. 160 p. URL: https://core.ac.uk/download/pdf/51213307.pdf
3. Shen C., Zhang K. Two-stage improved Grey Wolf optimization algorithm for feature selection on high-dimensional classification. *Complex & Intelligent Systems*. 2022;8(4):2769–2789. DOI: 10.1007/s40747–021–00452–4
4. Basak H., Das M., Modak S. RSO: A novel reinforced swarm optimization algorithm for feature selection. arXiv:2107.14199. URL: https://arxiv.org/pdf/2107.14199.pdf
5. Roffo G., Melzi S. Features selection via eigenvector centrality. In: Proc. 5th Int. workshop on new frontiers in mining complex patterns (NFMCP2016). (Riva del Garda, 19 September, 2016). Cham: Springer-Verlag; 2017. (Lecture Notes in Computer Science. Vol. 10312). URL: https://www.researchgate.net/publication/305918391_Feature_Selection_via_Eigenvector_Centrality
6. Hall M.A. Correlation-based feature selection for machine learning. PhD thesis. Hamilton: The University of Waikato; 1999. 198 p. URL: https://www.lri.fr/~pierres/donn%E 9es/save/these/articles/lpr-queue/hall99correlationbased.pdf
7. James G., Witten D., Hastie T., Tibshirani R. An introduction to statistical learning: With applications in R. 8th ed. New York, NY: Springer Science+Business Media; 2017. 440 p. (Springer Texts in Statistics).
8. Janitza S., Celik E., Boulesteix A.-L. A computationally fast variable importance test for random forests for high-dimensional data. *Advances in Data Analysis and Classification*. 2018;12(4):885–915. DOI: 10.1007/s11634–016–0276–4
9. Магнус Я.Р., Катышев П.К., Пересецкий А.А. Эконометрика. М.: Дело; 2004. 576 с.
10. Magnus Ya.R., Katyshev P.K., Peresetskii A.A. Econometrics. Moscow: Delo; 2004. 576 p. (In Russ.).
11. Pearson K. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*. 1901;2(11):559–572. DOI: 10.1080/14786440109462720
12. Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д. Прикладная статистика. Классификация и снижение размерности. М.: Финансы и статистика; 1989. 607 с.
13. Aivazyan S.A., Bukhshtaber V.M., Enyukov I.S., Meshalkin L.D. Applied statistics. Classification and dimensionality reduction. Moscow: Finansy i statistika; 1989. 607 p. (In Russ.).
14. Zhang Y., Dong Z., Phillips P., Wang S., Ji G., Yang J., Yuan T.-F. Detection of subjects and brain regions related to Alzheimer's disease using 3D MRI scans based on eigenbrain and machine learning. *Frontiers in Computational Neuroscience*. 2015;9:66. DOI: 10.3389/fncom.2015.00066
15. Hocking R.R. The analysis and selection of variables in linear regression. *Biometrics*. 1976;32(1):1–49. DOI: 10.2307/2529336
16. Mirjalili S., Mirjalili S.M., Lewis A. Grey wolf optimizer. *Advances in Engineering Software*. 2014;69:46–61. DOI: 10.1016/j.advengsoft.2013.12.007
17. Flom P.L., Cassell D.L. Stopping stepwise: Why stepwise and similar selection methods are bad, and what you should use. In: Northeast SAS Users Group 2007 (NESUG 2007). (Baltimore, 11–14 November, 2007). URL: https://www.lexjansen.com/pnwsug/2008/DavidCassell-StoppingStepwise.pdf
18. Eberhart R., Kennedy J. A new optimizer using particle swarm theory. In: Proc. 6th Int. symp. on micro machine and human science (MHS'95). (Nagoya, 04–06 October, 1995). Piscataway, NJ: IEEE; 1995:39–43. DOI: 10.1109/MHS.1995.494215

19. Schott J. R. Fault tolerant design using single and multicriteria genetic algorithm optimization. PhD thesis. Cambridge, MA: Massachusetts Institute of Technology; 1995. 201 p. URL: https://dspace.mit.edu/handle/1721.1/11582
20. Karaboga D. An idea based on honey bee swarm for numerical optimization. Technical Report. 2005;(06). URL: https://abc.erciyes.edu.tr/pub/tr06_2005.pdf
21. Altmann A., Toloşi L., Sander O., Lengauer T. Permutation importance: A corrected feature importance measure. *Bioinformatics*. 2010;26(10):1340–1347. DOI: 10.1093/bioinformatics/btq134
22. Hapfelmeier A., Ulm K. A new variable selection approach using random forests. *Computational Statistics & Data Analysis*. 2013;60:50–69. DOI: 10.1016/j.csda.2012.09.020
23. Louzada F., Ara A., Fernandes G. B. Classification methods applied to credit scoring: Systematic review and overall comparison. *Surveys in Operations Research and Management Science*. 2016;21(2):117–134. DOI: 10.1016/j.sorms.2016.10.001
24. Santosa F., Symes W. W. Linear inversion of band-limited reflection seismograms. *SIAM Journal on Scientific and Statistical Computing*. 1986;7(4):1307–1330. DOI: 10.1137/0907087
25. Hilt D. E., Seegrist D. W. Ridge: A computer program for calculating ridge regression estimates. USDA Forest Service Research Note. 1977;(236). URL: https://ia803007.us.archive.org/23/items/ridgecomputerpro236hilt/ridgecomputerpro236hilt.pdf
26. Тихонов А. Н. О решении некорректно поставленных задач и методе регуляризации. *Доклады Академии наук СССР*. 1963;151(3):501–504.
27. Tikhonov A. N. Solution of incorrectly formulated problems and the regularization method. *Soviet Mathematics. Doklady*. 1963;(4):1035–1038. (In Russ.: *Doklady Akademii nauk SSSR*. 1963;151(3):501–504.).
28. Воронцов К. В. Лекции по алгоритмам восстановления регрессии. 21 декабря 2007 г. URL: http://www.ccas.ru/voron/download/Regression.pdf
29. Vorontsov K. V. Lectures on regression recovery algorithms. December 21, 2007. URL: http://www.ccas.ru/voron/download/Regression.pdf (In Russ.).
30. Zou H., Hastie T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*. 2005;67(2):301–320. DOI: 10.1111/j.1467–9868.2005.00503.x
31. Брюс П., Брюс Э. Разведочный анализ данных. Практическая статистика для специалистов Data Science. Пер. с англ. СПб.: БХВ-Петербург; 2018:19–58.
32. Bruce P., Bruce A. Exploratory data analysis. In: Practical statistics for data scientists: 50 essential concepts. Beijing: O'Reilly Media; 2017;1–46. (Russ. ed.: Bruce P., Bruce A. Razvedochnyi analiz dannykh. Prakticheskaya statistika dlya spetsialistov Data Science. St. Petersburg: BHV-Peterburg; 2018:19–58.).
33. Afanasiev S., Smirnova A. Predictive fraud analytics: B-tests. *Journal of Operational Risk*. 2018;13(4):17–46. DOI: 10.21314/JOP.2018.213
34. Lin J. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*. 1991;37(1):145–151. DOI: 10.1109/18.61115
35. Kolmogorov A. Sulla determinazione empirica di una legge di distribuzione. *Giornale dell'Istituto Italiano degli Attuari*. 1933;4:83–91.
36. Harris D., Harris S. Digital design and computer architecture. 2nd ed. San Francisco, CA: Morgan Kaufmann; 2012. 720 p.
37. Kutner M. H., Nachtsheim C. J.; Neter J. Applied linear regression models. 4th ed. New York, NY: McGraw-Hill/Irwin; 2004. 701 p.

ИНФОРМАЦИЯ ОБ АВТОРАХ / ABOUT THE AUTHORS

Сергей Владимирович Афанасьев — магистрант, Национальный исследовательский университет «Высшая школа экономики», Москва, Россия; вице-президент, начальник управления статистического анализа КБ «Ренессанс Кредит», Москва, Россия

Sergey V. Afanasiev — Master's student, National Research University Higher School of Economics, Moscow, Russia; Vice President, Head of the Statistical Analysis Department, Bank "Renaissance Credit", Moscow, Russia

https://orcid.org/0000-0001-5119-507X

svafanasev@gmail.com

Диана Маратовна Котерева — магистрант, Национальный исследовательский университет «Высшая школа экономики», Москва, Россия; руководитель направления моделирования и оперативного анализа, КБ «Ренессанс Кредит», Москва, Россия
Diana M. Kotereva — Master's student, National Research University Higher School of Economics, Moscow, Russia; Head of Modeling and Operational Analysis Department Bank "Renaissance Credit", Moscow, Russia
https://orcid.org/0000-0001-6102-0222
dmkotereva@edu.hse.ru

Алексей Алексеевич Мироненков — старший преподаватель кафедры эконометрики и математических методов экономики Московской школы экономики, Московский государственный университет имени М.В. Ломоносова, Москва, Россия
Alexey A. Mironenkov — Senior Lecturer at the Department of Econometrics and Mathematical Methods of Economics Moscow School of Economics, Lomonosov Moscow State University, Moscow, Russia
https://orcid.org/0000-0001-5754-8825
Автор для корреспонденции / Corresponding author:
mironenkov@mse-msu.ru

Анастасия Андреевна Смирнова — магистрант, Национальный исследовательский университет «Высшая школа экономики», Москва, Россия; начальник отдела разработки и анализа эффективности скоринговых систем, КБ «Ренессанс Кредит», Москва, Россия
Anastasiya A. Smirnova — Master's student, National Research University Higher School of Economics, Moscow, Russia; Head of Scoring Systems Department Bank "Renaissance Credit", Moscow, Russia
https://orcid.org/0000-0002-1836-1555
aasmirnova_24@edu.hse.ru

Заявленный вклад авторов:
С.В. Афанасьев — постановка проблемы, разработка концепции статьи, критический анализ литературы, описание результатов и формирование выводов исследования.
Д.М. Котерева — разработка методологии, программного кода, сбор статистических данных, табличное и графическое представление результатов.
А.А. Мироненков — описание результатов и формирование выводов исследования.
А.А. Смирнова — критический анализ литературы, разработка программного кода.

***Authors' declared contribution*:**
S.V. Afanasiev — problem statement, conceptualisation of the article, critical analysis of the literature, description of the results and formation of the conclusions of the study.
D.M. Kotereva — development of methodology, program code, statistical data collection, tabular and graphical presentation of results.
A.A. Mironenkov — description of the results and formation of the conclusions of the study.
A.A. Smirnova — critical analysis of the literature, development of the program code.

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.
Conflicts of Interest Statement: The authors have no conflicts of interest to declare.

Статья поступила в редакцию 11.02.2022; после рецензирования 25.02.2022; принята к публикации 27.12.2022.
Авторы прочитали и одобрили окончательный вариант рукописи.
The article was submitted on 11.02.2022; revised on 25.02.2022 and accepted for publication on 27.12.2022.
The authors read and approved the final version of the manuscript.