Сергей АФАНАСЬЕВ и др.

Двухлесовый метод отбора признаков по сравнению с регрессионными методами показывает лучшее качество для нелинейных моделей и сопоставимое качество для линейных. При этом двухлесовый метод работает в десятки раз быстрее, что является важным преимуществом для решения современных практических задач моделирования, где используются большие высокоразмерные выборки. В статье описана методика комбинированного отбора признаков с использованием двухлесового метода, которая была разработана в банке «Ренессанс Кредит» и используется для разработки моделей.

Сергей АФАНАСЬЕВ, *КБ «Ренессанс Кредит» (ООО), исполнительный директор, начальник управления статистического анализа*
Диана КОТЕРЕВА, *КБ «Ренессанс Кредит» (ООО), руководитель направления моделирования и оперативного анализа*
Анастасия СМИРНОВА, *КБ «Ренессанс Кредит» (ООО), начальник отдела разработки и анализа эффективности скоринговых систем*

Two-forest jump: комбинированный отбор признаков с использованием двухлесового метода

Обзор методов отбора признаков

Методы машинного обучения успешно применяются в розничном кредитовании и хорошо зарекомендовали себя в таких направлениях, как оценка кредитоспособности заемщика (scoring), взыскание просроченной задолженности (collection) и перекрестные продажи (CRM). За годы применения моделей в розничном кредитовании банки накопили огромные массивы данных, содержащие разностороннюю информацию о заемщиках. С одной стороны, это позволило более точно предсказывать поведение заемщика, с другой — породило проблему избыточности данных, которая сильно усложняет разработку моделей. Для решения этой проблемы предлагаются различные методы отбора признаков, которые позволяют повысить качество моделей и упростить их разработку[1].

[1] Guyon I., Elisseeff A. An Introduction to Variable and Feature Selection // Journal of Machine Learning Research. 2003. Vol. 3. P. 1157-1182.

Two-forest jump: комбинированный отбор признаков с использованием двухлесового метода

Предлагаемые в научных исследованиях методы отбора признаков делятся на три типа[1]:

1) *фильтры (Filter methods)* — выбирают переменные независимо от модели и базируются на общих характеристиках, таких как корреляция с целевой переменной. Среди наиболее известных фильтров можно выделить матрицу корреляций;

2) *обертки (Wrapper methods)* — оценивают поднаборы переменных и позволяют обнаружить возможную взаимосвязь между переменными. К этим методам относятся такие популярные алгоритмы, как Forward Selection, Backward Elimination, RFE и др.;

3) *вложения (Embedded methods)* — встроены в алгоритмы обучения и осуществляют отбор признаков в процессе построения модели. К популярным вложенным методам относятся LASSO, гребневая регрессия и др.

Среди классических подходов наибольшую популярность в прикладных задачах получили:

— матрица корреляций (Correlation Feature Selection, CFS);

— метод главных компонент (Principal Component Analysis, PCA);

— ступенчатая регрессия: метод прямого отбора (Forward Selection), метод обратного исключения (Backward Elimination) и метод последовательного отбора (Stepwise).

Среди продвинутых (современных) подходов на практике используются:

— отбор с использованием алгоритмов случайного леса и градиентного бустинга;

— встроенные методы: LASSO (L1-регуляризация), гребневая регрессия (L2-регуляризация), ElasticNet и др.

Матрица корреляций

Отбор переменных с помощью матрицы корреляций 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[2]. Данный подход получил широкое распространение в банковской сфере и до недавнего времени являлся отраслевым стандартом.

К плюсам метода можно отнести:

— простоту реализации;

— простоту интерпретации.

[1] Hamon J. Optimisation combinatoire pour la sélection de variables en régression en grande dimension: Application en génétique animale. Université des Sciences et Technologie de Lille – Lille I, 2013.

[2] Hall M.A. Correlation-based Feature Selection for Machine Learning. The University of Waikato, 1999.

Среди недостатков метода можно выделить следующие:
— метод чувствителен к качеству данных (выбросы, ошибки);
— метод не учитывает сложные взаимосвязи;
— могут быть ошибки при интерпретации.

Метод главных компонент

Метод был предложен Карлом Пирсоном в 1901 г.[1], но до сих пор популярен в прикладных задачах машинного обучения. Суть метода состоит в уменьшении размерности признакового пространства путем вычисления главных компонент матрицы признаков и последующем понижении размерности матрицы через ее сингулярное разложение[2].

Среди плюсов метода можно отметить простоту реализации, поэтому он часто используется при первичном отборе признаков.

Из недостатков метода выделяют следующие[3]:

— метод не учитывает целевую переменную, из-за чего главные компоненты могут оказаться не самыми информативными;

— метод чувствителен к масштабу;

— могут возникнуть проблемы с выбором порога отсечения для главных компонент.

Ступенчатая регрессия

Методы ступенчатой регрессии относятся к методам-оберткам, суть которых заключается в построении моделей на разных наборах признаков и последующем сравнении результатов моделей на тестовой выборке. Среди методов-оберток наибольшую популярность в банковских задачах получили методы ступенчатой регрессии:

— метод прямого отбора (Forward Selection)[4];

— метод обратного исключения (Backward Elimination)[5];

— метод последовательного отбора (Stepwise)[6].

Несмотря на свою простоту, методы Forward Selection, Backward Elimination и Stepwise устарели и подвергались неоднократной критике[7].

[1] Pearson K. On Lines and Planes of Closest Fit to Systems of Points in Space // Philosophical Magazine. 1901. Vol. 2. No. 11. P. 559–572.

[2] Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д. Прикладная статистика. Классификация и снижение размерности. М.: Финансы и статистика, 1989.

[3] Воронцов К.В. Введение в машинное обучение. НИУ ВШЭ & Yandex School of Data Analysis (coursera.org).

[4] Efroymson M.A. Multiple regression analysis. In: Mathematical Methods for Digital Computers. Ralston A. and Wilf H.S. (eds.). New York: Wiley, 1960.

[5] Паклин Н., Орешков В. Бизнес-аналитика: от данных к знаниям. СПб.: Питер, 2009.

[6] SAS/STAT User's Guide. Version 6. Fourth Edition. Volume 2. Cary, NC: SAS Institute Inc., 1989.

[7] Flom P.L., Cassell D.L. Stopping stepwise: Why stepwise and similar selection methods are bad, and what you should use. NorthEast SAS Users Group, 2007.

Two-forest jump: комбинированный отбор признаков с использованием двухлесового метода

Поэтому в прикладных задачах все большую популярность набирают методы, основанные на алгоритмах случайного леса и градиентного бустинга.

Случайный лес и градиентный бустинг

Подходы, основанные на алгоритмах случайного леса и градиентного бустинга, относятся к методам-оберткам. Данные подходы можно отнести к современным методам, поскольку они являются предметом последних научных исследований, а также используются на популярных соревновательных площадках, таких как Kaggle.

Основным плюсом данных методов является их высокая точность. К важным преимуществам методов, основанных на алгоритмах случайного леса, относят возможность получения несмещенных оценок важностей признаков за счет использования полностью рандомизированных деревьев, что позволяет улучшить качество итоговых моделей[1].

Из минусов данных методов можно отметить их вычислительную сложность и возможные негативные эффекты от переобучения.

Главным плюсом методов, основанных на алгоритмах случайного леса и градиентного бустинга, является их высокая точность.

Методы регуляризации

Регуляризация относится к группе вложенных методов, в которых отбор признаков становится частью процесса построения модели. Общая концепция регуляризации заключается в добавлении штрафного слагаемого к оптимизируемому функционалу ошибки, которое наказывает модель за чрезмерную сложность. В логистической регрессии, ставшей банковским стандартом, наиболее распространенными методами регуляризации являются L1 и L2. Регуляризация L1 позволяет обнулять часть весовых коэффициентов регрессии, а регуляризация L2 ограничивает норму весовых коэффициентов регрессии[2]. Таким образом регуляризация помогает решить проблему роста дисперсии при избыточном количестве признаков в модели, поскольку ограничение на норму весов ограничивает рост дисперсии.

Существует множество других методов отбора признаков, которые применяются на практике индивидуально или в комбинации[3]. В этой статье мы рассмотрим комбинированный отбор признаков

[1] Parr T., Turgutlu K., Csiszar C., Howard J. Beware Default Random Forest Importances. March 26, 2018 (explained.ai/rf-importance/).

[2] Воронцов К.В. Лекции по алгоритмам восстановления регрессии. 21 декабря 2007 г.

[3] Hamon J. Optimisation combinatoire pour la sélection de variables en régression en grande dimension: Application en génétique animale. Université des Sciences et Technologie de Lille — Lille I, 2013.

с использованием двухлесового метода и покажем преимущества данного подхода по сравнению с классическими методами отбора.

Комбинированный отбор признаков

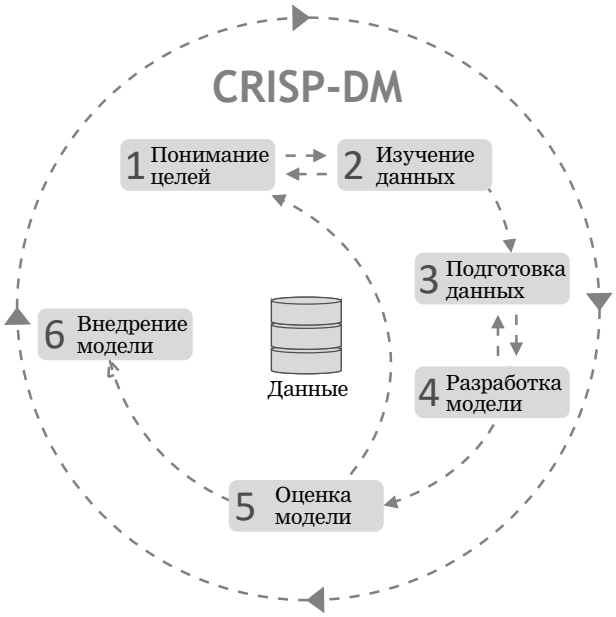
Процесс разработки и внедрения моделей в банке можно разбить на шесть фаз, установленных межотраслевым стандартом исследования данных CRISP-DM (рис. 1)[1]:

1) понимание бизнес-целей задачи;
2) изучение доступных данных;
3) подготовка данных;
4) разработка модели;
5) оценка качества модели;
6) внедрение модели.

Подготовка данных и разработка модели — центральные фазы процесса моделирования. Именно на этих этапах проводится отбор

Рисунок 1

Схема межотраслевого стандарта разработки моделей CRISP-DM[2]



[1] Shearer C. The CRISP-DM model: the new blueprint for data mining // Journal of Data Warehousing. 2000. Vol. 5. P. 13-22.

[2] Последовательность фаз не определена строго. В большинстве проектов приходится возвращаться к предыдущим этапам, а затем снова двигаться вперед.

Two-forest jump: комбинированный отбор признаков с использованием двухлесового метода

признаков. Согласно последним научным исследованиям, наиболее эффективными техниками отбора признаков являются методы, построенные на алгоритмах случайного леса. Возникает вопрос: если в научных исследованиях показано, что какой-то метод эффективнее другого, почему не использовать именно его? Наиболее точный ответ на этот вопрос был сформулирован в высказывании, которое приписывают американскому бейсболисту Йоги Берре: «В теории нет разницы между теорией и практикой. А на практике есть».

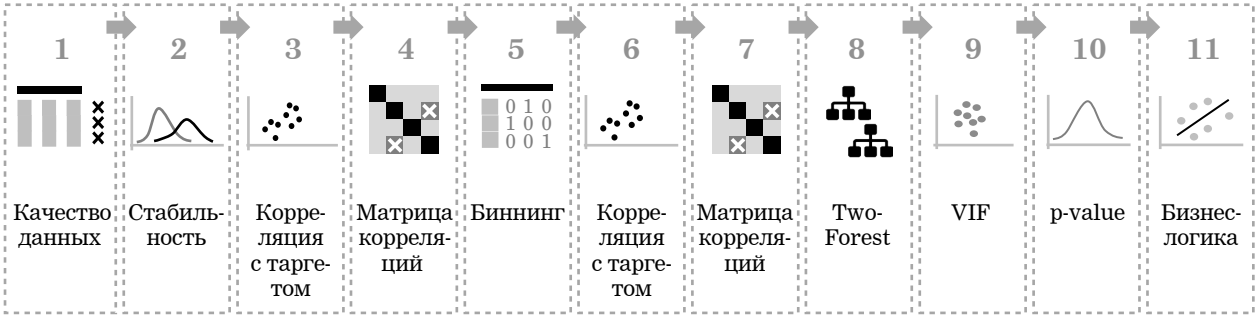
Если говорить о методах отбора, предлагаемых в научных исследованиях, то можно увидеть ряд недостатков применительно к практическим задачам. Как правило, исследования проводятся на открытых базах данных, которые содержат либо мало наблюдений, либо мало признаков. Например, открытые медицинские базы данных могут содержать информацию всего о нескольких десятках или сотнях пациентов, а доступные базы данных кредитных заявок содержат всего несколько десятков признаков.

При этом выборки, которые мы используем в банке для разработки моделей, содержат одновременно и большое количество признаков (от нескольких сотен до нескольких тысяч), и большое количество наблюдений (от 100 тысяч до нескольких миллионов). На таких выборках методы, предлагаемые в научных исследованиях, либо не дают декларируемого результата, либо работают крайне долго из-за высокой ресурсоемкости.

Именно эти проблемы мотивировали нас использовать комбинированную схему отбора признаков, которая включает в себя 11 шагов (рис. 2). Каждый из этих шагов мы подробно разберем далее.

Рисунок 2

Схема комбинированного отбора признаков



Сергей АФАНАСЬЕВ и др.

Шаг 1. Анализ качества данных

Качество данных — обобщенное понятие, отражающее степень пригодности данных к решению определенной задачи. В соответствии с международными и российскими стандартами основными критериями качества данных являются точность (accuracy), полнота (completeness), согласованность (consistency), достоверность (credibility), правильность (correctness), доступность (availability) и др.[1] Оценка и мониторинг качества данных проводятся на разных этапах жизненного цикла модели:

1) на стадии формирования и сбора данных;
2) на стадии разработки модели;
3) на стадии эксплуатации модели.

На стадии разработки модели действует принцип GIGO[2] («Garbage In, Garbage Out» — «Мусор на входе, мусор на выходе»), который означает, что при неверных входящих данных будут получены неверные результаты, даже если алгоритм правилен. Таким образом, необходимым условием высокого качества модели является высокое качество данных, на которых разрабатывается модель. Поэтому контроль качества данных может быть сведен к контролю качества модели. Однако оценка качества данных на стадии моделирования может способствовать улучшению качества и стабильности модели, поэтому оценка качества данных и действия по его повышению на стадии разработки моделей являются важным этапом аналитического проекта.

Оценка качества данных на стадии моделирования может способствовать улучшению качества и стабильности модели. Методы анализа качества данных: разведочный анализ, анализ пропусков и неполноты данных, анализ аномалий и др.

Среди основных проблем, вызывающих снижение качества данных, обычно выделяют следующие[3]:

— пропуски или неполнота данных;
— орфографические ошибки;
— аномалии;
— фиктивные значения;
— логические несоответствия;
— закодированные значения;
— несоответствие форматов;
— дублирование;
— ложность информации;
— противоречивость информации;

[1] См. стандарты по качеству данных ISO 9000:2015, ISO/TS 8000, ISO/IEC 25012:2008 и ГОСТ Р 8000.

[2] Work With New Electronic «Brains» Opens Field For Army Math Experts // The Hammond Times. Retrieved March 20, 2016. P. 65 (via Newspapers.com).

[3] Полянский Ю.Н. Основы оценки кредитного риска по методологии Базель II: Семинар-практикум / Институт банковского дела Ассоциации российских банков, 2020.

Two-forest jump: комбинированный отбор признаков с использованием двухлесового метода

— избыточность информации;

— и др.

Среди методов анализа качества данных можно выделить:

1) разведочный анализ (Exploratory Data Analysis, EDA)[1] — 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 (например, логарифмическая), расчет *z*-score, выделение «хвостов» через IQR[2] и др.

Шаг 2. Проверка стабильности и непрерывности данных

Большинство статистических методов для разработки моделей предполагают использование непрерывных и стабильных данных. Причиной нестабильности в данных может быть изменение бизнес-процессов банка, законодательства, клиентского потока или клиентского поведения, форматов данных и др. Поэтому одним из важных этапов работы с данными является проверка условия непрерывности и стабильности. Далее описан метод оценки стабильности признаков, разработанный в банке «Ренессанс Кредит» и применяемый при разработке банковских моделей.

Перед оценкой стабильности все строковые переменные необходимо преобразовать в числовой формат с помощью метода LabelEncoder[3] (пропуски заменяются уникальным числовым значением). После предобработки данных проводится оценка стабильности пере-

[1] Брюс П., Брюс Э. Разведочный анализ данных // Практическая статистика для специалистов Data Science. СПб.: БХВ-Петербург, 2018. С. 19–58.

[2] Upton G.J.G., Cook I.T. Understanding Statistics. Oxford University Press, 1996.

[3] scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html.

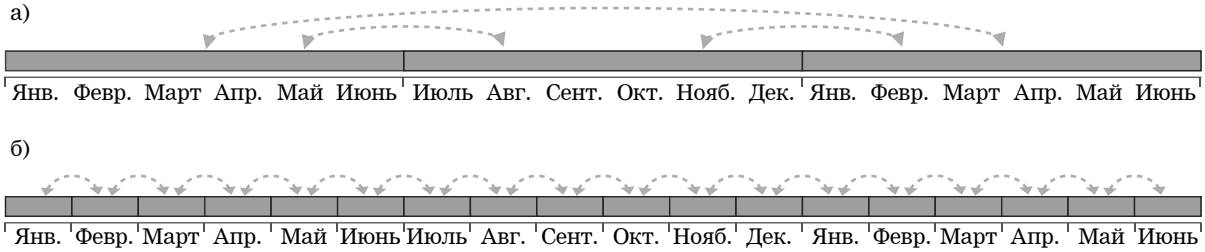
менных. Для этого вычисляются статистики *S*, *PSI* и *KS* для распределений, построенных на разных периодах. Оценка стабильности проводится на периодах двух размеров:

1) *большие периоды*. Общая выборка делится на равные подвыборки с большим интервалом (например, по полугодиям[1]), после чего на этих подвыборках попарно сравниваются распределения признаков по принципу «каждый с каждым» (рис. 3а);

2) *маленькие периоды*. Общая выборка делится на равные подвыборки с маленьким интервалом (например, по месяцам), после чего попарно сравниваются распределения признаков по смежным периодам (рис. 3б).

Рисунок 3

Пример разбиения выборки на периоды для оценки стабильности переменной: а) большие периоды; б) маленькие периоды



Статистики *S* и *PSI* рассчитываются на данных плотностей распределений, построенных на двух подвыборках за разные периоды одинаковой длины.

S-статистика рассчитывается по формуле:

$$S = \sum_{i=1}^{n} \frac{|a_i - b_i|}{2}, \quad (1)$$

где i — порядковый номер интервала разбиения;

a_i — значения плотности распределения на выборке A в период T_1;

b_i — значения плотности распределения на выборке B в период T_2.

Значения *S*-статистики нормированы и лежат в диапазоне [0; 1].

Статистика PSI рассчитывается по формуле:

$$PSI = \sum_{i=1}^{n} (a_i - b_i) \ln\left(\frac{a_i}{b_i}\right). \quad (2)$$

[1] Размеры больших периодов либо выбираются аналитиком и зависят от типа задачи, либо настраиваются как гиперпараметры модели и подбираются автоматически.

Two-forest jump: комбинированный отбор признаков с использованием двухлесового метода

Значения статистики *PSI* лежат в диапазоне [0; ∞).

Статистика KS рассчитывается на данных функций распределений (кумулятивных плотностей), построенных на двух выборках за одинаковый период:

$$KS = \max_i(|F(a_i) - F(b_i)|), \quad (3)$$

где $F(a_i)$ — значения функции распределения на выборке A в период T_1;

$F(b_i)$ — значения функции распределения на выборке B в период T_2.

Значения статистики *KS* лежат в диапазоне [0; 1]. Вычисление *S*-статистики и статистики *KS* проиллюстрировано рис. 4.

После расчета статистик *S*, *PSI* и *KS* применяется процедура их усреднения и категоризации по алгоритму, представленному в табл. 1.

Рисунок 4

Иллюстрация вычисления S-статистики и статистики Колмогорова-Смирнова (KS)

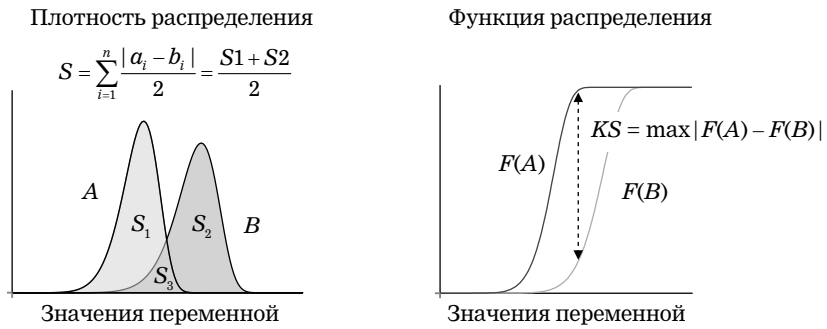


Таблица 1

Пример расчета пороговых значений статистик и весов для отбора переменных

Статистика	Стабильна	Слегка нестабильна	Нестабильна
KS	[0%; 10%)	[10%; 25%)	[25%; 100%]
S	[0%; 10%)	[10%; 25%)	[25%; 100%]
PSI	[0; 10)	[10; 25)	[25; ∞)
Присвоенный вес	**0**	**0,5**	**1**
Большие периоды: Big = max(KS_i) + max(S_i) + max(PSI_i)	[0; 1]	[1; 2)	[2; 3)
Маленькие периоды: Small = Avg(KS_j + S_j + PSI_j)	[0; 1]	[1; 1,5)	[1,5; 3)
Присвоенный вес	**0**	**0,5**	**1**
Максимум (консервативный подход): Max(Big; Small)	{0}	{0,5}	{1}
Усреднение (лояльный подход): 0,5Big + 0,5Small	[0; 0,2)	[0,2; 0,6)	[0,6; 1]
Присвоенный вес	**0**	**0,5**	**1**

После всех вычислений нестабильные переменные удаляются, слегка нестабильные и стабильные переменные остаются в выборке для дальнейшего анализа.

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

Шаг 3. Корреляция признаков с целевой переменной

Анализ корреляции признаков с целевой переменной позволяет отобрать признаки, сильно влияющие на целевую переменную. Данный метод не учитывает сложные зависимости между признаками, поэтому его можно отнести к «грубым» методам, которые можно применять для первичного отбора, когда расширенный список признаков очень большой. Методы анализа корреляции зависят от типа целевой переменной и типа исследуемого признака[1].

1. Для бинарной целевой переменной:

— для непрерывных признаков рассчитываются статистика Стьюдента (если признак распределен нормально) или тест Манна–Уитни (если признак распределен ненормально). Проверка на нормальность проводится с помощью теста Колмогорова–Смирнова;

— для категориальных и бинарных признаков рассчитывается хи-квадрат (критерий Пирсона).

2. Для категориальной целевой переменной (количество категорий больше 2):

— для непрерывных признаков проводится тест ANOVA;

— для категориальных и бинарных признаков рассчитывается хи-квадрат (критерий Пирсона).

3. Для непрерывной целевой переменной:

— для непрерывных признаков рассчитывается корреляция Пирсона;

— для категориальных признаков проводится тест ANOVA;

— для бинарных признаков рассчитываются статистика Стьюдента (если признак распределен нормально) или тест Манна–Уитни (если признак распределен ненормально).

Признаки, которые не проходят тест на значимость корреляции с целевой переменной, исключаются из дальнейшего анализа.

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

Two-forest jump: комбинированный отбор признаков с использованием двухлесового метода

Шаг 4. Матрица корреляций

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

$$R_x = \begin{pmatrix} 1 & r_{x_1x_2} & \dots & r_{x_1x_n} \\ r_{x_2x_1} & 1 & \dots & r_{x_2x_n} \\ \dots & \dots & \dots & \dots \\ r_{x_nx_1} & r_{x_nx_2} & \dots & 1 \end{pmatrix}, \quad (4)$$

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. Матрицы для непрерывных и категориальных признаков строятся и анализируются отдельно, поскольку сравнивать корреляцию Пирсона и корреляцию Спирмена не совсем корректно.

После построения матрицы корреляций необходимо исключить из дальнейшего анализа высококоррелированные признаки. Чтобы из двух высококоррелированных признаков выбрать наименее значимый, необходимо рассчитать корреляции всех признаков с целевой переменной. Для разных типов признаков и целевых переменных рассчитываются разные коэффициенты корреляций:

1. Для бинарной целевой переменной:

— для непрерывных признаков рассчитывается корреляция Пирсона;

— для категориальных и бинарных признаков рассчитывается фи-статистика (формула (5)).

2. Для категориальной целевой переменной (количество категорий больше 2):

— для непрерывных признаков проводится тест ANOVA;

— для категориальных и бинарных признаков рассчитывается фи-статистика.

3. Для непрерывной целевой переменной:

— для непрерывных признаков рассчитывается корреляция Пирсона;

— для категориальных признаков проводится тест ANOVA;

— для бинарных признаков рассчитывается корреляция Пирсона.

Фи-статистика рассчитывается по формуле:

$$\varphi = \sqrt{\frac{Chi}{n}}, \quad (5)$$

где Chi — значение коэффициента хи-квадрат;

n — количество наблюдений в выборке.

Матрица корреляций хорошо подходит для первичной фильтрации признаков перед многофакторным анализом, в котором используются продвинутые методы.

Фи-статистика может принимать максимальное значение 0,707 (при $Chi = 1$ и $n = 2$).

Отбор признаков проводится по следующему алгоритму:

1) если один из признаков в паре принадлежит к «слабо нестабильным», то признак автоматически удаляется из дальнейшего анализа;

2) если оба признака принадлежат либо к «слегка нестабильным», либо к «стабильным», то удаляется тот признак, который имеет меньшее значение корреляции с целевой переменной;

3) если признаки имеют одинаковую стабильность и равные значения корреляций с целевой переменной, то удаляется любой из двух признаков.

Отбор признаков с помощью матрицы корреляции выполняется два раза — до бинаризации признаков и после. Пороговые значения для отбора признаков задаются аналитиком и зависят от типа задачи. По умолчанию рекомендуется брать пороги значений корреляции 90% на этапе «до бинаризации признаков» (слабая фильтрация) и 70% на этапе «после бинаризации признаков» (сильная фильтрация).

Важно отметить, что корреляция Спирмена построена на рангах, поэтому результаты оценки с ее помощью могут оказаться некорректными для категориальных непорядковых признаков. Поэтому при первом отборе не рекомендуется ставить низкий порог.

Использование матрицы корреляций при отборе признаков имеет как преимущества, так и недостатки, о которых мы говорили в начале статьи. Поэтому матрица корреляций хорошо подходит для первичной фильтрации признаков перед многофакторным анализом, в котором используются продвинутые методы.

Шаг 5. Dummy-кодирование категориальных переменных

После отбора признаков с помощью «грубых» методов необходимо преобразовать категориальные переменные в бинарные[1]. Для логистической регрессии используется процедура *dummy-кодирования*[2] по методу полного ранга: одна из категорий удаляется (если в данных есть пропуски, то удаляется бин с пропусками). То есть после dummy-кодирования категориальной переменной получается $k - 1$ новых бинарных переменных, где k — количество категорий в исходной переменной.

[1] Это необходимое условие для использования алгоритмов логистической регрессии – переменные должны быть числовыми или бинарными (не категориальными).

[2] Другие названия метода dummy-кодирования – one-hot encoding и биннинг.

Two-forest jump: комбинированный отбор признаков с использованием двухлесового метода

Важно отметить, что мы применяем процедуру dummy-кодирования только к категориальным признакам. Бинаризацию числовых признаков мы не делаем по нескольким причинам:

— во-первых, бинаризация числовых признаков увеличивает размерность признакового пространства — а это то, с чем мы боремся, применяя различные методы отбора;

— во-вторых, в научном сообществе критикуются классические методы бинаризации числовых признаков, а в качестве альтернативы предлагается использовать сплайны (к сплайнам мы вернемся в экспериментальной части нашей статьи)[1];

— в-третьих, при наличии в выборке большого количества признаков (а у нас выборки высокоразмерные) практически для любого нелинейного числового признака с высокой вероятностью найдутся категориальные признаки, которые его бинаризуют. Например, категориальная переменная «семейное положение» (холост, женат/замужем, вдовец/вдова) может очень хорошо бинаризовать переменную «возраст». То есть работа, которую проводят аналитики, разрабатывая много признаков, по сути уже является экспертной бинаризацией числовых признаков.

Шаг 6. Корреляция признаков с целевой переменной после биннинга

После бинаризации необходимо провести повторный анализ корреляций признаков с целевой переменной и удалить признаки, слабо коррелированные с целевой переменной. Данный этап идентичен шагу 3.

Шаг 7. Матрица корреляций после биннинга

После бинаризации признаков также необходимо провести повторный анализ совместных корреляций, поскольку полученные бинарные признаки могут сильно коррелировать друг с другом или с другими признаками.

Алгоритм отбора идентичен шагу 4, однако в силу специфики показателя корреляции Спирмена на данном шаге рекомендуется выставлять более низкий порог корреляций для отсеивания признаков. Порог выбирается аналитиком и зависит от специфики данных. По умолчанию рекомендуется выставлять порог корреляции, равный 70%.

[1] Bennette C., Vickers A. Against quantiles: categorization of continuous variables in epidemiologic research, and its discontents // BMC Medical Research Methodology. 2012. Vol. 12 (doi.org/10.1186/1471-2288-12-21).

Шаг 8. Двухлесовый (two-forest) метод

После первичной фильтрации признаков с помощью корреляционных методов необходимо провести отбор переменных более точными методами, учитывающими сложные взаимосвязи между признаками. К таким методам относятся алгоритмы, построенные на методе случайного леса (random forest)[1].

Для отбора переменных методом случайного леса необходимо оценить важность (importance) каждой переменной. Существует два основных подхода для оценки важности:

1. *Важность на основе уменьшения неоднородности:*

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

b) полученная на шаге (a) сумма делится на общее количество деревьев;

c) шаги (a) и (b) повторяются для всех переменных.

Важность — это частота использования переменной в качестве предиктора ветвления.

Двухлесовый метод помогает решить проблему скорости и точности отбора признаков для высокоразмерных выборок.

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

a) обучается модель случайного леса;

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

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

d) по новой выборке рассчитывается ошибка;

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

Вычисленная разность ошибок является пермутированной важностью переменной.

Практически все методы отбора переменных, построенные на алгоритме случайного леса, чувствительны к размерности признакового пространства и требуют больших вычислительных ресурсов. Именно поэтому перед их применением требуется проводить предварительный отбор признаков с помощью более простых подходов.

Решить проблему скорости и точности на высокоразмерных выборках попытались исследователи из Мюнхенского университета, пред-

[1] Breiman L. Random forests // Machine Learning. 2001. Vol. 45. P. 5–32 (doi:10.1023/A:1010933404324).

[2] 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).

Two-forest jump: комбинированный отбор признаков с использованием двухлесового метода

ложив двухлесовый метод для отбора признаков[1]. Далее опишем концепцию двухлесового метода и его модернизацию, которую мы используем в нашем банке[2].

Общая идея двухлесового метода заключается в оценке важности признаков как качества прогнозирования при случайной перестановке (пермутации):

$$VI_j = P\left(Y \neq f\left(X_1,...,X_j^*,...,X_p\right)\right) - P\left(Y \neq f\left(X_1,...,X_j,...,X_p\right)\right). \quad (6)$$

Двухлесовый метод работает по следующему алгоритму:

1) исходная выборка делится на две подвыборки;

2) на каждой из двух подвыборок строится случайный лес;

3) для построенных моделей для каждой *j*-й переменной вычисляется важность при перестановке VI_j на множестве, которое не использовалось для построения модели;

4) определяются множества:

M1 = {все отрицательные важности};

M2 = {все нулевые важности};

M3 = {все отрицательные важности × (–1)};

5) на множестве M = M1 U M2 U M3 строится функция плотности распределения *F* (рис. 5);

6) для каждой *j*-й переменной рассчитывается значение *p*-value = $= 1 - F(VI_j)$.

Адаптированный двухлесовый метод работает по следующему алгоритму:

1) обучающая выборка стратифицированно разбивается на две равные части;

2) на каждой из двух подвыборок обучается случайный лес[3];

3) на второй отложенной подвыборке оценивается качество модели;

4) каждая переменная случайным образом пермутируется и считается результат для каждой из двух моделей на отложенных подвыборках;

5) рассчитывается разница между baseline-значением (шаг 3) и новым значением;

6) важность переменной рассчитывается как среднее значение важностей по двум подвыборкам;

[1] Janitza S., Celik E., Boulesteix A.-L. A computationally fast variable importance test for random forests for high-dimensional data. Springer-Verlag Berlin Heidelberg, 2016 (DOI 10.1007/s11634-016-0270-x).

[2] В оригинальной статье метод называется New Approach. Но поскольку многие исследователи считают свои подходы новыми и часто называют их New Approach, чтобы не было путаницы с другими New Approach, мы назвали метод Two-Forest (двухлесовый), поскольку обучаются сразу два случайных леса.

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

Рисунок 5

Пример построения распределения F на основе переменных, которые могут быть нерелевантными (с отрицательными или нулевыми оценками важности)

7) рассчитывается значение *p*-value для важности:

— выбираются наблюдения с отрицательными значениями средней важности;

— выбираются наблюдения с нулевыми значениями средней важности;

— отрицательные значения важности умножаются на (–1);

— векторы, полученные на шагах 1–3, конкатенируются;

— для полученного вектора строится эмпирическое кумулятивное распределение;

— на построенном эмпирическом распределении рассчитывается *p*-value;

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

— разделить важность на среднее значение baseline. Выбираются те переменные, у которых изменение больше заданного порога (порог подбирается аналитиком);

— отсортировать переменные по значениям важности и выбрать первые *N* переменных (число *N* подбирается аналитиком). При этом следует учитывать, что значение *p*-value по выбранным переменным должно быть меньше 10%.

Преимущества двухлесового метода:

1) метод имеет высокое качество, сопоставимое с качеством других методов случайного леса;

2) метод работает значительно быстрее стандартных методов случайного леса;

3) метод может работать с данными большой размерности (> 1000 переменных), а на данных меньшей размерности (~100 переменных) сохраняет значение ошибки 1-го рода.

Two-forest jump: комбинированный отбор признаков с использованием двухлесового метода

Из недостатков двухлесового метода можно выделить то, что:

1) метод может плохо работать с маленьким количеством незначимых переменных;

2) при взаимной корреляции двух сильных признаков могут быть удалены оба признака;

3) на выборках большой размерности метод может показывать низкую статистическую силу (когда самая сильная переменная может пройти отбор с вероятностью всего 15–20%).

Шаг 9. Проверка мультиколлинеарности (VIF)

Другой подход для снижения мультиколлинеарности между признаками основан на оценке показателя VIF (Variance Inflation Factor), с целью расчета которого строятся линейные регрессии для каждого объясняющего признака (выступающего в качестве целевой переменной) от всех остальных признаков.

Отбор признаков с помощью показателя VIF осуществляется по следующему алгоритму:

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

$$X_i = \beta_0 + \sum_{j=1}^{k} \beta_j X_j,\ i \neq j, \tag{7}$$

где β_0 — свободный член регрессии;

k — общее количество признаков (включая анализируемый).

2. Рассчитывается коэффициент VIF для признака X_i по формуле:

$$\mathrm{VIF}_t = \frac{1}{1 - R_i^2}, \tag{8}$$

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

3. Проводится оценка полученных значений VIF, где применяется общее эмпирическое правило: признаки со значением VIF > 10 относятся к мультиколлинеарным[1].

Из списка всех мультиколлинеарных признаков удаляется признак с максимальным значением VIF.

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

Для проверки статистической значимости признаков можно использовать один из тестов: тест отношения правдоподобия, тест Вальда и тест множителей Лагранжа.

[1] Kutner M.H., Nachtsheim C.J., Neter J. Applied Linear Regression Models, 4th ed. Chicago: McGraw-Hill/Irwin, 2004.

Сергей АФАНАСЬЕВ и др.

Шаг 10. Проверка статистической значимости признаков

Среди тестов для проверки статистической значимости признаков можно выделить тест отношения правдоподобия, тест Вальда и тест множителей Лагранжа. Все три теста асимптотически эквивалентны, хотя для конечных выборок могут не совпадать[1]. Поэтому для проверки значимости можно использовать один из этих тестов, например тест 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 — совокупность (вектор) некоторых функций параметров. Для проверки нулевой гипотезы сравниваются функции правдоподобия полной модели (обученной на n признаках) и укороченной модели без тестируемого признака (обученной на $n-1$ признаках). Для этого рассчитывается статистика отношения правдоподобия (likelihood ratio test, LR):

$$\mathrm{LR} = 2(L_l - L_s) = 2\mathrm{In}\frac{L_l}{L_s}, \quad (9)$$

где L_l — значение логарифмической функции правдоподобия полной модели;

L_s — значение логарифмической функции правдоподобия укороченной модели.

Несмотря на то что регуляризация сама является встроенным методом для отбора признаков, при хорошо отлаженной схеме предварительного отбора признаков регуляризация почти не влияет на финальное качество модели.

Статистика LR при нулевой гипотезе имеет распределения хи-квадрат с q степенями свободы — $\chi^2(q)$, где q — количество ограничений (исключенных признаков). Если значение данной статистики больше критического значения распределения при заданном уровне значимости, то исключенный признак считается значимым и предпочтение отдается полной модели. В противном случае исключенная переменная признается незначимой.

Шаг 11. Экспертный анализ

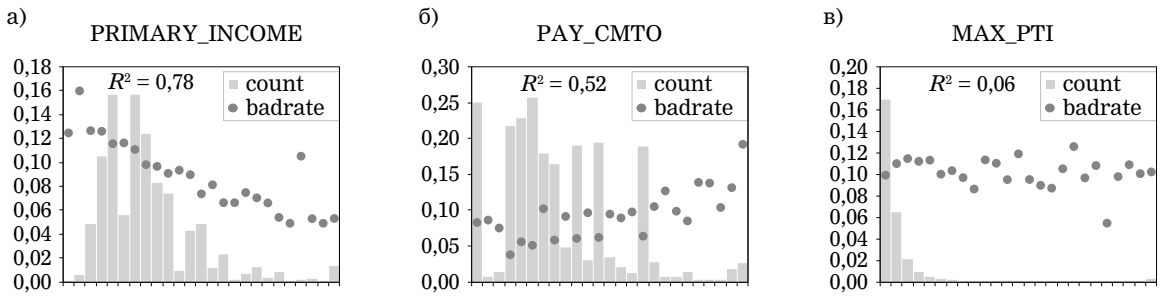
Для контроля соответствия отобранных признаков бизнес-смыслу необходимо проводить экспертный анализ, который сводится к графической интерпретации зависимостей между отобранными признаками и целевой переменной, а также соответствия знака весового коэффициента признака в регрессионной модели знаку корреляции (рис. 6).

[1] Магнус Я.Р., Катышев П.К., Пересецкий А.А. Эконометрика. М.: Дело, 2004.

Two-forest jump: комбинированный отбор признаков с использованием двухлесового метода

Рисунок 6

Пример графического экспертного анализа признаков: а) сильная отрицательная линейная зависимость с целевой переменной; б) умеренная положительная линейная зависимость с целевой переменной; в) слабая линейная зависимость с целевой переменной



Важно отметить, что экспертный анализ является субъективным методом и зависит от опыта аналитика, поэтому лучше использовать его в качестве метода контроля результатов, полученных на шагах 1–10.

Эксперименты

Для сравнения подходов мы выбрали три метода отбора признаков, два из которых популярны в банковской практике:

1) *Matrix* — отбор с использованием матрицы корреляций и однофакторных логрегрессий;

2) *Forward* — отбор с использованием метода прямого отбора;

3) *Two-Forest* — отбор с использованием двухлесового метода.

Чтобы не сравнивать слабые и сильные методы, мы оставили почти все шаги комбинированной схемы отбора, изменив только этапы со сравниваемыми методами. Также мы убрали из схемы последний шаг с экспертным анализом, чтобы исключить субъективную составляющую.

Сравнение методов проводилось на четырех банковских моделях:

1) *CRM: PTB (probability to buy*) — оценка вероятности отклика клиента на кросс-сейл предложение;

2) *Scoring: Application PD* — оценка вероятности дефолта на этапе выдачи кредита;

3) *Scoring: Behavioral PD* — оценка вероятности дефолта в течение жизни кредита с использованием поведенческой информации о предыдущих платежах;

4) *Collection: Allocation* — оценка вероятности «переката» просрочки в более поздний бакет.

В качестве алгоритмов моделирования использовались два метода:

1) логистическая регрессия с L2-регуляризацией;

2) градиентный бустинг в реализации LightGBM.

Для логистической регрессии проводилась стандартизация признаков. С помощью кросс-валидации (5-fold) подбирались гиперпараметры class_weight (балансировка классов) и коэффициент регуляризации C. Несмотря на то что регуляризация сама является встроенным методом для отбора признаков, при хорошо отлаженной схеме предварительного отбора признаков регуляризация почти не влияет на финальное качество модели. Тем не менее, мы оставляем этот гиперпараметр для дополнительного контроля[1].

Для алгоритма LightGBM использовался фиксированный набор гиперпараметров (бенчмарк): boosting_type='gbdt', max_depth=3, num_leaves=31, learning_rate=0.1, feature_fraction=0.9, bagging_fraction=0.9, bagging_freq=5, n_estimators=100, class_weight={0: 2*w_b, 1: 1}, где параметр w_b — это дисбаланс классов.

Также в общей схеме отбора в качестве гиперпараметра подбирался коэффициент корреляции для матрицы корреляций после биннинга (подбор по двум значениям — 50% и 70%).

Для разработки моделей использовалась обучающая выборка (train), для оценки качества модели — валидационная выборка (test) и out-off-time выборка (OOT). Для оценки качества моделей и методов отбора использовалась метрика Gini[2].

Эффективность методов отбора

Результаты сравнения трех схем представлены в табл. 2 и на рис. 7.

Схемы Forward и Two-Forest в среднем по всем бизнес-моделям и ML-алгоритмам показывают качество лучше, чем схема Matrix.

Сравнение схем Forward и Two-Forest показывает, что схема Forward работает на логистической регрессии чуть лучше схемы Two-Forest. При этом для алгоритма LightGBM схема Two-Forest дает лучшее качество моделей, чем схема Forward, в среднем на 0,5% Gini.

Незначительный проигрыш схемы Two-Forest схеме Forward на логистической регрессии можно объяснить тем, что Two-Forest хорошо отбирает нелинейные признаки, на которых линейные клас-

[1] Если в процессе подбора гиперпараметров выясняется, что регуляризатор значительно влияет на качество модели, то схема отбора признаков плохо настроена или работает неправильно.

[2] При тестировании использовалась также метрика Average Precision, но мы оставили только Gini, чтобы упростить читаемость таблиц. По метрике Average Precision получились аналогичные результаты.

Two-forest jump: комбинированный отбор признаков с использованием двухлесового метода

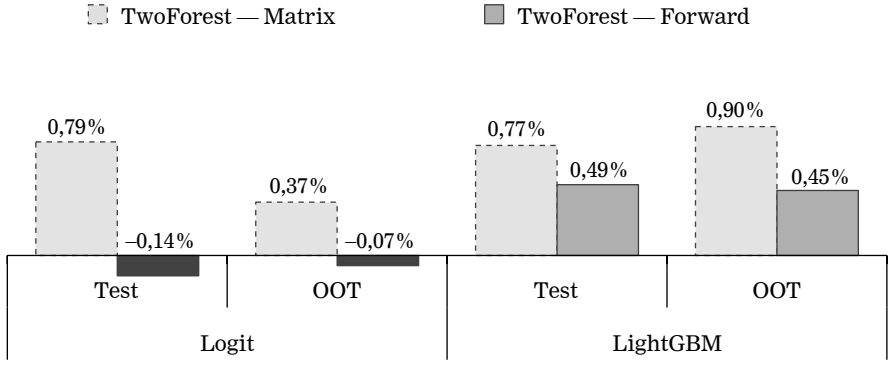
Таблица 2

Сравнение методов отбора признаков (%)

Алгоритм	Модель	GINI (TEST)			GINI (OOT)			Delta GINI (TEST)			Delta GINI (OOT)		
		Matrix	Forward	2Forest	Matrix	Forward	2Forest	Forward — Matrix	2Forest — Matrix	2Forest — Forward	Forward — Matrix	2Forest — Matrix	2Forest — Forward
LogReg	PTB (CRM)	41,57	42,59	43,32	43,12	43,69	43,40	1,02	1,74	0,73	0,57	0,28	–0,30
	Behavioral PD (Scoring)	68,43	69,07	68,81	64,00	64,66	64,39	0,64	0,38	–0,25	0,67	0,40	–0,27
	Application PD (Scoring)	40,51	41,64	40,93	39,80	40,67	40,51	1,13	0,42	–0,72	0,87	0,71	–0,16
	Allocation (Collection)	60,48	61,43	61,11	60,75	60,41	60,85	0,95	0,62	–0,32	–0,34	0,11	0,44
	Среднее по моделям							**0,93**	**0,79**	**–0,14**	**0,44**	**0,37**	**–0,07**
LGBM	PTB (CRM)	43,80	43,02	44,01	44,80	44,81	45,27	–0,79	0,21	1,00	0,01	0,47	0,46
	Behavioral PD (Scoring)	69,04	70,68	70,50	64,93	67,05	66,82	1,64	1,46	-0,18	2,12	1,89	–0,23
	Application PD (Scoring)	42,51	43,56	43,90	41,79	42,03	42,90	1,05	1,39	0,34	0,24	1,11	0,87
	Allocation (Collection)	64,99	64,18	65,00	64,94	64,36	65,07	–0,81	0,01	0,82	–0,58	0,13	0,71
	Среднее по моделям							**0,27**	**0,77**	**0,49**	**0,45**	**0,90**	**0,45**

Рисунок 7

Усредненная разница финального качества моделей между разными схемами отбора признаков (разница по метрике Gini)



сификаторы работают хуже, чем нелинейные (LGBM и др.). И наоборот, в методе Forward используется линейная регрессия, то есть данный метод лучше отбирает линейные признаки, на которых хорошо работает логистическая регрессия. При этом полученные результаты показывают, что LightGBM дает лучшее качество, чем логистическая регрессия, в среднем на 2,2% Gini. Решить проблему нелинейности признаков и повысить качество моделей логистической регрессии можно с помощью сплайнов, результаты по которым будут показаны далее.

Если сравнивать методы Forward и Two-Forest по времени работы (табл. 3), то Two-Forest работает в десятки раз быстрее. В данном эксперименте сравнивалось время работы самих методов, а не полных схем отбора.

Таблица 3

Сравнение методов по времени работы

Модель	Кол-во наблюдений (Train)	Исходное кол-во признаков	Время (чч:мм:сс)		
			Forward	2Forest	Forward/ 2Forest
PTB (CRM)	303 220	1 222	14:01:28	0:30:48	74
Behavioral PD (Scoring)	588 385	1 087	16:11:04	0:16:36	58
Application PD (Scoring)	497 063	423	5:34:12	0:15:00	22
Allocation (Collection)	172 250	162	3:06:40	0:05:50	32

Схему отбора признаков можно представить в виде пошаговой воронки, в которой отражается количество оставшихся и удаленных признаков на каждом этапе отбора.

Мы видим, что по сравнению с классическими подходами двухлесовый метод дает сразу два улучшения на больших выборках:

1) статистически значимое улучшение финального качества моделей;

2) существенное преимущество в скорости работы.

На этом можно было бы подвести черту, но есть еще несколько деталей в комбинированной схеме отбора, на которые хотелось бы обратить внимание.

Воронка отбора признаков

Схему отбора признаков можно представить в виде пошаговой воронки, в которой отражается количество оставшихся и удаленных признаков на каждом этапе отбора (табл. 4). Полученная воронка позволяет проанализировать, какой метод дал наибольший вклад при отборе, а также сделать выводы о качестве данных и корректности работы методов. Мы разберем пример анализа воронок для

Two-forest jump: комбинированный отбор признаков с использованием двухлесового метода

Таблица 4

Воронки поэтапного отбора признаков для модели Behavioral PD

Схема Matrix	Кол-во признаков	Удалено/добавлено	Схема Forward	Кол-во признаков	Удалено/добавлено	Схема Two-Forest	Кол-во признаков	Удалено/добавлено
	1 087			1 087			1 087	
1. Качество данных	722	–365	1. Качество данных	722	–365	1. Качество данных	722	–365
2. Стабильность	321	–401	2. Стабильность	321	–401	2. Стабильность	321	–401
3. Таргет-корреляция	320	–1	3. Таргет-корреляция	320	–1	3. Таргет-корреляция	320	–1
4. Матрица	170	–150	4. Матрица	170	–150	4. Матрица	170	–150
5. Биннинг	238	68	5. Биннинг	238	68	5. Биннинг	238	68
6. Таргет-корреляция	237	–1	6. Таргет-корреляция	237	–1	6. Таргет-корреляция	237	–1
7. Матрица	111	–126	7. Матрица	111	–126	7. Матрица	111	–126
8. Gini	70	–41	8. Forward	54	–57	8. Two-Forest	70	–41
9. VIF	70	0	9. VIF	53	–1	9. VIF	64	–6
10. *p*-value	52	–18	10. *p*-value	45	–8	10. *p*-value	55	–9
Итого отобрано:	**52**		**Итого отобрано:**	**45**		**Итого отобрано:**	**55**	

модели Behavioral PD в рамках трех схем отбора: Matrix, Forward и Two-Forest. Для других моделей анализ воронки проводится аналогично.

Шаг 1. Проверяется качество данных. На этом этапе было удалено 34% от всех поданных признаков: это высокий показатель, что может указывать на низкое качество данных. В задаче разработки модели Behavioral PD высокий процент признаков низкого качества был связан с большим окном выборки для разработки модели, из-за чего большая часть переменных, построенная на свежих источниках данных, не заполнялась для старых кредитных заявок. Если для неглубоких выборок эта доля является высокой, то следует провести аудит признаков: возможно, часть из них работает некорректно.

Шаг 2. Проверяется стабильность признаков. По воронке видно, что было удалено 56% признаков от всех поданных на этот шаг. Данный показатель также является высоким, но связано это с большой глубиной выборки: новые сконструированные признаки имели

много пропусков на всей глубине выборки. Если на неглубоких выборках получается много нестабильных признаков, необходимо проанализировать данные и параметры настройки алгоритма. Иногда периоды для проверки стабильности хорошо работают на одних задачах и плохо на других, поэтому необходима дополнительная настройка алгоритма. В нашей практике мы сталкивались с такой проблемой в collection-моделях, когда на поздних стадиях взыскания признаки обновляются редко и для анализа стабильности необходимо укрупнение временных интервалов.

Шаг 3. Проверяется корреляция с целевой переменной и отсеиваются слабокоррелированные признаки. На данном этапе из 321 признака был удален всего один, что является адекватным показателем. Если исключается слишком много признаков, то необходимо провести аудит удаленных переменных: возможно, что они не подходят для поставленной задачи.

Комбинированная схема отбора, используемая в нашем банке, включает в себя проверку качества и стабильности признаков, одномерные методы-фильтры и многомерные методы-обертки.

Шаг 4. Проводится отбор с помощью корреляционной матрицы. Доля удаленных признаков составила 47%. Это хороший показатель: при большом количестве признаков матрица корреляций должна удалять примерно 30–50% переменных. Если матрица удаляет мало признаков, то либо на данный этап попало мало признаков (менее 100), либо выставлены очень высокие пороги корреляций и их надо пересмотреть.

Шаг 5. Проводится бинаризация категориальных переменных, поэтому общее количество признаков увеличивается и необходимо повторно пройти шаги 3 и 4 (соответственно здесь они будут шагами 6 и 7), чтобы понизить размерность признакового пространства.

Шаг 8. Сравниваются три метода: однофакторные логистические регрессии (отбор признаков с Gini > 5%), метод Forward Selection и двухлесовый метод. Поскольку на этом этапе применяется главный метод-обертка в общей схеме, на данном шаге должно быть удалено значительное количество признаков (от 30 до 70%). Если удаляется мало признаков, то либо метод-обертка плохо настроен, либо на предыдущих шагах использовались слишком жесткие методы-фильтры, отсекающие много признаков.

Шаг 9. Проводится дополнительная проверка мультиколлинеарности признаков с помощью показателя VIF, то есть данный шаг является контролем для предыдущих этапов. Если на данном шаге исключается большая доля признаков, то, значит, на предыдущих шагах плохо отработали матрица корреляций и (или) двухлесовый метод.

Шаг 10. По оставшимся признакам оценивается значение *p*-value с помощью отношения правдоподобий полной регрессии и регрессии

Two-forest jump: комбинированный отбор признаков с использованием двухлесового метода

без анализируемого признака. То есть этот метод является упрощенной версией метода обратного исключения (Backward Elimination). Поэтому данный этап отбора по аналогии с шагом 9 является контролем предыдущих шагов и количество удаленных признаков на этом этапе должно быть небольшим.

Сплайны для числовых признаков

Алгоритм LightGBM продемонстрировал существенный отрыв в качестве от логистической регрессии по всем исследуемым задачам (см. табл. 2). При этом двухлесовый метод показывает лучшее качество именно при использовании LightGBM. Однако ансамблевые методы не всегда возможно применить на практике, особенно в задачах скоринга, где регулятор может потребовать интерпретируемости результатов.

Экспериментальным путем мы определили, что на моделях логистической регрессии двухлесовый метод отбора признаков не дает преимущества в качестве по сравнению с регрессионными методами. Это связано с тем, что двухлесовый метод умеет отбирать сильные нелинейные признаки, с которыми логистическая регрессия работает плохо. Решить проблему можно методами энкодинга переменных, включая критикуемый Фрэнком Хареллом биннинг и предлагаемую им альтернативу — сплайны[1].

В табл. 5 показаны результаты применения сплайнов к числовым признакам для модели предсказания дохода клиента. Видно, что сплайны улучшают качество линейной регрессии по метрике R^2, хотя сплайн-регрессия все еще отстает от LightGBM.

Таблица 5

Сравнение алгоритмов моделирования для предсказания дохода клиента (%)

Модель (сегмент)	Алгоритм	R2			R2-adjusted		
		TRAIN	TEST	OOT	TRAIN	TEST	OOT
Income (CASH_top_up)	LinearRegression	54,9	55,0	50,0	54,9	55,0	50,0
	Splines	57,2	57,3	53,1	57,2	57,3	53,1
	LGBM	60,6	59,0	55,4	60,6	59,0	55,4
Income (CASH_non_top_up)	LinearRegression	46,7	46,6	45,6	46,7	46,6	45,6
	Splines	49,0	49,1	48,5	49,0	49,1	48,5
	LGBM	51,5	51,0	50,3	51,5	51,0	50,3

[1] biostat.mc.vanderbilt.edu/wiki/Main/CatContinuous.

Заключение

В 1995 г. американский математик Дэвид Волперт сформулировал свою знаменитую теорему «о бесплатных завтраках», в которой утверждается, что самый изощренный алгоритм, который мы можем придумать, в среднем по всевозможным задачам дает такое же качество, как простейшее предсказание — все точки принадлежат одному классу[1].

И хотя позже было показано, что теорема справедлива только при определенных условиях, которые редко встречаются в реальной жизни[2], в научных исследованиях можно увидеть, что пока нет одного алгоритма, который показывал бы лучшее качество на всех исследуемых задачах[3].

Схожие мысли высказывал Константин Воронцов на конференции Scoring Days 2018: «С точки зрения академической науки, когда я начинал заниматься машинным обучением 25 лет назад, тот научный коллектив, в который я пришел еще студентом, в общем-то жил с полной уверенностью (и она была основана на примерно 30-летнем опыте предыдущих исследований), что задачу можно решать любым методом».

Перенося идеи Волперта и Воронцова на задачу отбора признаков, можно предположить, что отбор можно делать любым методом. Но за 25 лет в машинном обучении кое-что изменилось:

— появились компьютерные мощности для работы с большими данными;

— появились новые алгоритмы для работы с большими выборками;

— раскрыли свой потенциал и некоторые старые алгоритмы, которые ранее не показывали преимущества на маленьких выборках.

Пермутация переменных, лежащая в основе двухлесового метода, становится не только предметом научных исследований, но и инструментом для прикладных задач. В соревновательных задачах Kaggle стали встречаться решения с использованием пермутации для отбора признаков. В популярную библиотеку Eli5 недавно был добавлен класс eli5.sklearn.permutation_importance[4].

[1] Wolpert D.H., Macready W.G. No Free Lunch Theorems for Optimization // IEEE Transactions on Evolutionary Computation. 1997. Vol. 1. Issue 1. P. 67–82 (CiteSeerX 10.1.1.138.6606; doi:10.1109/4235.585893).

[2] Streeter M. Two Broad Classes of Functions for Which a No Free Lunch Result Does Not Hold // Genetic and Evolutionary Computation – GECCO 2003. P. 1418–1430.

[3] Lessmann S. et al. Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research // European Journal of Operational Research. 2015. Vol. 247. Issue 1. P. 124–136 (DOI: 10.1016/j.ejor.2015.05.030).

[4] eli5.readthedocs.io/en/latest/blackbox/permutation_importance.html.

машинное обучение \ случайный лес \ анализ качества данных

Two-forest jump: комбинированный отбор признаков с использованием двухлесового метода

Комбинированная схема отбора, используемая в нашем банке, включает в себя проверку качества и стабильности признаков, одномерные методы-фильтры и многомерные методы-обертки.

Проверка качества и стабильности переменных позволяет использовать один большой универсальный набор признаков для разных типов задач.

Одномерные методы-фильтры позволяют отсеивать мусорные признаки, а методы-обертки являются тонкой настройкой и учитывают сложные многомерные зависимости.

Включение в комбинированную схему отбора нескольких методов-оберток позволяет контролировать корректность работы методов на предыдущих шагах.

Таким же контролирующим свойством обладают и встроенные методы регуляризации. Если регуляризация значимо улучшает качество модели, то, скорее всего, комбинированная схема отбора плохо настроена или в используемых данных есть ошибки, которые нужно исправить.

С помощью воронки можно анализировать работу каждого шага комбинированной схемы отбора. Это позволяет находить ошибки в данных и обнаруживать некорректно настроенные методы для определенных типов задач.

Комбинированную схему отбора признаков можно полностью автоматизировать, встроив ее в общий pipeline моделирования с последующей разработкой моделей в режиме End2End. Это позволяет ускорить процесс разработки моделей и снизить модельные риски (ошибки экспертного анализа). Пороговые значения метрик, а также порядок методов в комбинированной схеме отбора можно дополнительно настраивать как гиперпараметры.