Publishing House "ANALITIKA RODIS" (analitikarodis@yandex.ru) http://publishing-vak.ru/

УДК 33

Вероятность оттока клиента при реализации скоринговой модели в условиях деятельности природохозяйственного предприятия

Лютягин Дмитрий Владимирович

Кандидат экономических наук, доцент,
Российский государственный геологоразведочный университет,
117485, Российская Федерация, Москва, ул. Миклухо-Маклая, 23;
e-mail: l-d-v@list.ru

Забайкин Юрий Васильевич

Кандидат экономических наук, доцент,
Российский государственный геологоразведочный университет,
117485, Российская Федерация, Москва, ул. Миклухо-Маклая, 23;
e-mail: 89264154444@yandex.ru

Аннотация

В работе показано применение различных моделей оценки клиентов природохозяйственного предприятия. В частности показано, что кроме того, изучением кредитной истории занимается банковский работник, в результате на принимаемое решение очень сильно воздействует субъективный фактор, а неточность финансовых показателей, возникающая по разным причинам, часто приводит к ошибкам и возникновению рискованных ситуаций. Поэтому существует необходимость в наукоемких механизмах, способных в определенном смысле заменить кредитных экспертов, сократить время анализа заявки и одновременно уменьшить долю субъективизма в принятии решения. В качестве такого механизма может быть использована скоринговая модель. При построении скоринговой модели возникает две основные проблемы. Первой является определение характеристик, которые нужно включать в построение модели и которые должны быть наиболее тесно связаны с ненадежностью или надежностью клиента. Данные характеристики должны содержать в себе необходимый объем информации, с помощью которого можно будет их классифицировать. С данными характеристиками связана одна из основных проблем скоринга, которая заключается в динамичности развития социальных и экономических процессов, с течением времени люди, обстоятельства, условия, могут измениться. Поэтому скоринговые модели необходимо разрабатывать на выборке из наиболее «свежих» клиентов, периодически проверять качество работы системы и при ухудшении обновлять или создавать новую модель. Для сельскохозяйственных организаций данная проблема также определяется тем, что поддержка кредитными средствами требуется постоянно.

Для цитирования в научных исследованиях

Лютягин Д.В., Забайкин Ю.В. Вероятность оттока клиента при реализации скоринговой модели в условиях деятельности природохозяйственного предприятия // Экономика: вчера, сегодня, завтра. 2019. Том 9. № 5В. С. 543-550.

Ключевые слова
Модель, клиент, переменные, логический шаг, операция.

Введение

Для форирования модели скоринга в учетах различных видов отраслевых риском необходимо осуществлять цензорирование и генерацию независимых переменных и целевого поля. Цензурирование происходило формированием целевого поля по признаку того, вернется ли клиент после определенного фиксированного дня операции [Kim, Seongkon, Won, and Don, 2014].

Далее происходило построение целевого поля по следующей схеме:

– если клиент в день наблюдения проводил операцию, то это наблюдение учитывается в выборке для вклада;

– если клиент в день наблюдения делал операцию, то это наблюдение учитывается в выборке для банка.

Таким образом одно наблюдение может попасть в обеих выборок.

Следующий логический шаг – это формирование «меры отсутствия» как перерыв в днях между двумя последовательными наблюдениями. При этом пропуск считается значимым, если его мера отсутствия удовлетворяет 3 условиям [Carminati, 2014]:

– если клиент отсутствовал более 3 дней;

– если значение меры отсутствие является экстремальным, а именно попадает в 0,975 квартиль распределения всех мер отсутствия, что наблюдались у определенного клиента;

– максимальный перерыв между операциями составляет не менее 15 дней.

Основная часть

Для построения моделей был сгенерирован ряд лаговых переменных, то есть значение наблюдений с лагом в один день. Если клиент отсутствовал в этот день, то осуществлялась линейная аппроксимация между двумя ближайшими наблюдениями его деятельности [Hand, 2008].

Происходила построение моделей на основе этого целевого поля и входных данных с помощью 2 типов моделей.

Для этого каждая выборка (для банка и клиента) поделилась на 2 части: тренировочную и тестовую. Разделение происходил случайным образом, но так, чтобы тренировочная покрывала 80% всех данных. В качестве метрики оценки использовался AUC, а целевой функцией выступала невязка логистической регрессии [Molloy, 2017].

За базовую модель для сравнения была взята обычная логистическая регрессия. Ее построение происходила с использованием формул (1-2), что имплементированы в стандартной библиотеке языка программирования R.

Тест Стьюдента на значимость коэффициентов регрессии показал, что большинство регрессоров мало коррелируют с целевым полем.

Результат логистической регрессии (рисунок 1).

Для улучшения результатов было решено взять еще большую глубину лага переменных, а именно: одну неделю. Однако это не дало ожидаемых результатов.

Для построения бустингового ансамбля деревьев использовалась библиотека xgboost из официального репозитория [Adewumi, Aderemi, and Andronicu, 2017].

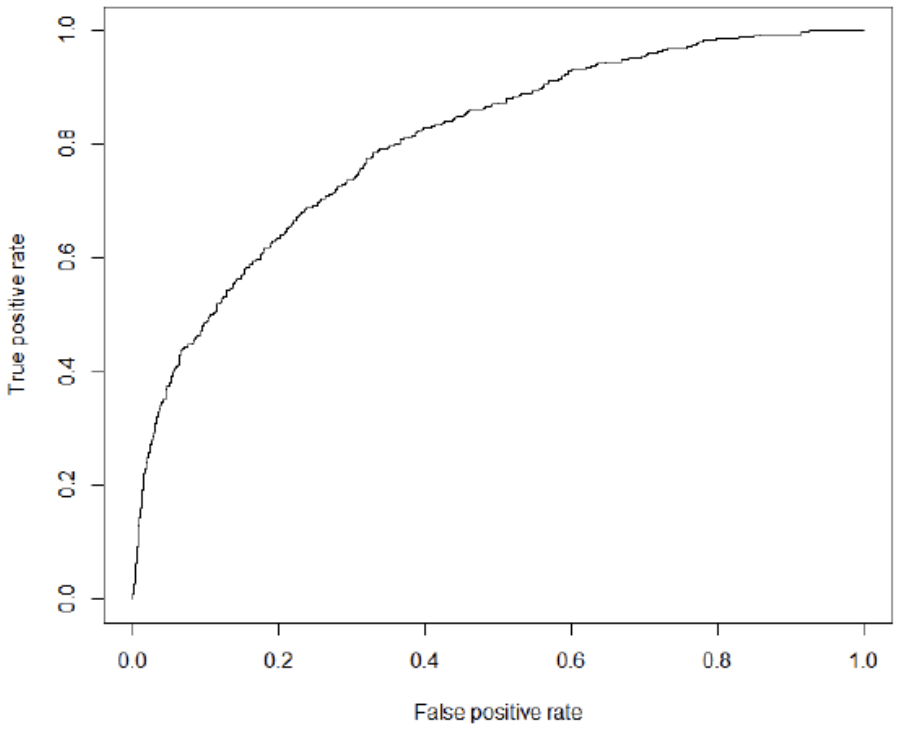

Рисунок 1 – ROC-кривая логистической регрессии

Для подбора оптимальных значений параметров модели (тюнинг) осуществлялся эмпирический поиск на сетке (max.depth; eta; nrounds) с шагами (1; 0,01; 1). См. рисунки 2-5.

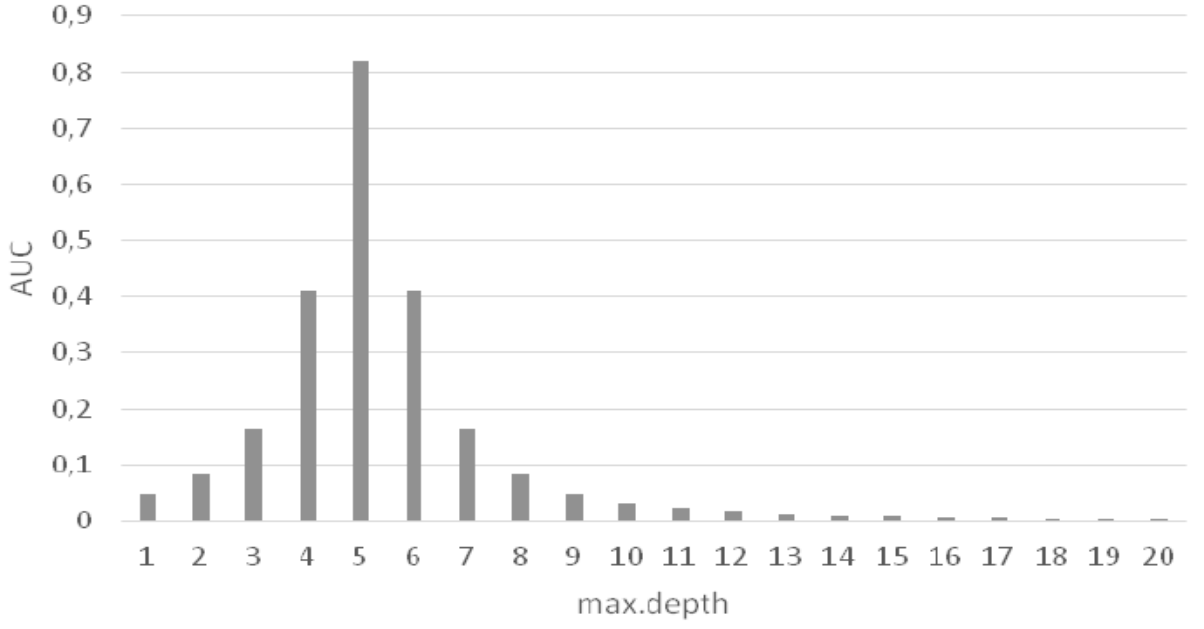

Рисунок 2 – Зависимость AUC от параметра max.depth (покер)

Зависимость между показателем AUC и параметром max.depth имеет ярко выраженный экстремум в точке 5. Такое значение этого параметра и взято за остаточный, поскольку оно максимизирует желаемую точность.

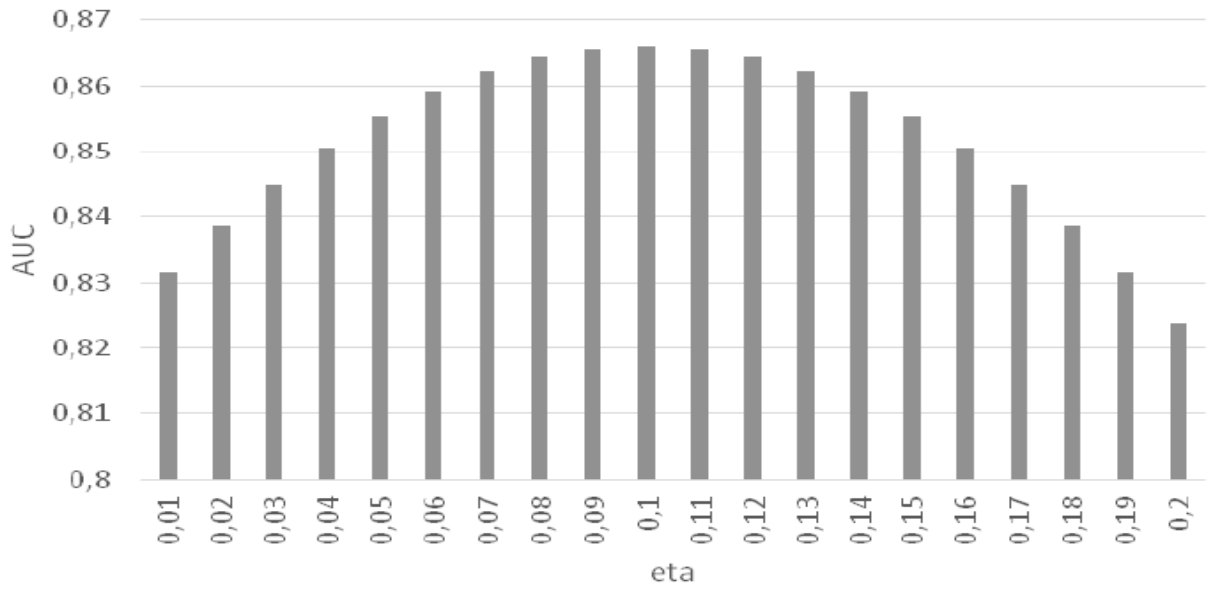

Рисунок 3 – Зависимость AUC от параметра eta

Зависимость между показателем AUC и параметром eta имеет распределение, похожее на нормальное. Максимум достигается в районе точки 0.1.

Сравнение регрессоров происходило по средневзвешенному приросту информации (см. рисунок 3).

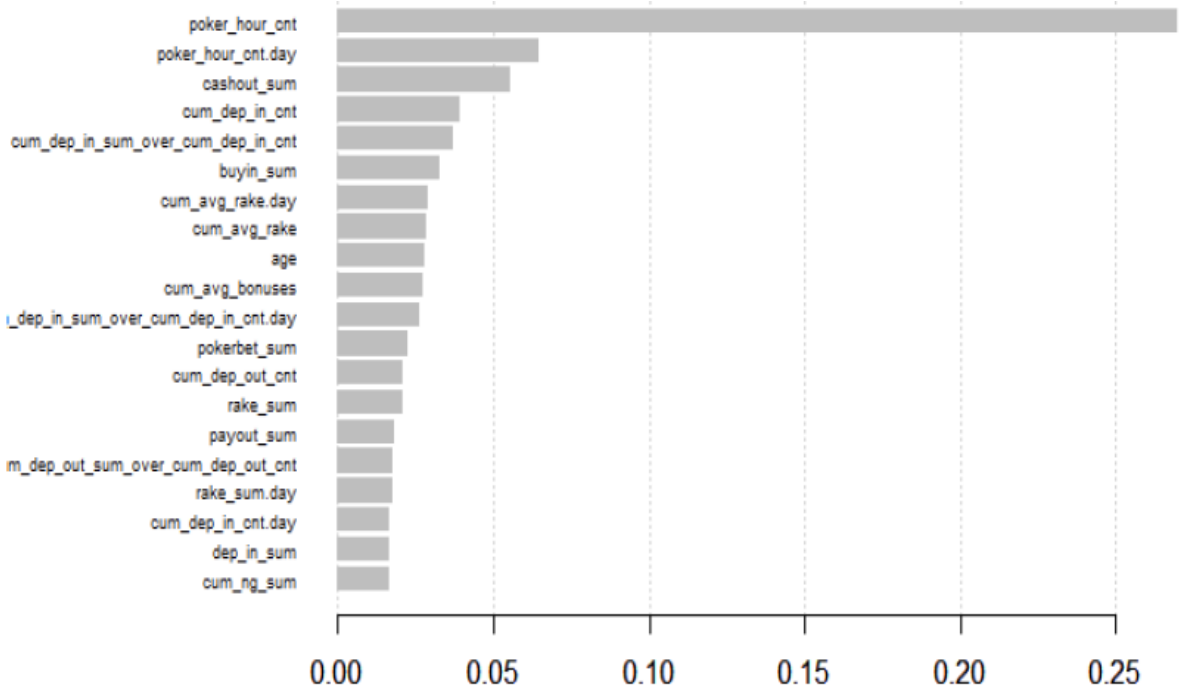

Рисунок 4 – График важности регрессоров

Наиболее важным показателем для модели оттока клиентов является количество совершенных операций в течение дня. Также важнми являются следующие показатели:

– количество проведенных операций в течение предыдущего дня;
– сумма выводов;
– кумулятивное количество депозитов;
– кумулятивное среднее значение величины депозита;
– сумма денежных взносов;
– среднее значение кумулятивного рейка за предыдущий день и др.

Учитывая список наиболее важных параметров, можно сказать, что на вероятность дефолта влияют не только показатели, касающиеся операций клиента, но также и его финансовые показатели [Khattri, Vipin, and Deepak, 2018].

Самым важным, как и в случае модели для кредиторов, является количество свершенных операций в течение дня. Однако в противовес операциям для клиентов МФО также важен показатель их доходности, то есть сколько они провели кумулятивно в течение всей своей деятельности как клиент банка, а также возраст клиента [Blackwell, Clive, 2014]. Стоит также отметить, что к одним из наиболее важных показателей попали и финансовые показатели, такие как:

– сумма операций;
– средняя сумма депозита;
– количество выводов;
– и что характерно средний кумулятивный рейк.

Одним из результатов этой работы является интерпретация полученных результатов модели см. формулу (1). Поскольку в модели оценки вероятности оттока не учитывается то, сколько дней клиент уже был отсутствующим, через значительную корреляцию с целевым полем, то для формирования окончательного балла используется преобразования в виде функции, принимающей на вход два параметра: вероятность, полученную ансамблем деревьев и фактическое количество дней, что клиент отсутствовал:

$$s = \frac{1}{1 + e^{\frac{-(n-\mu)}{\sigma}}} \tag{1}$$

Где:

$$\mu = m \frac{\ln \frac{1 - s_0}{s_0}}{\ln \frac{(1 - s_0) s_1}{s_0 (1 - s_1)}}$$

$$\sigma = \frac{m}{\ln \frac{(1 - s_0) s_1}{s_0 (1 - s_1)}}$$

n – количество дней отсутствия на момент расчета;

$$S_1(S_0) = \begin{cases} 0.95 & S_0 < 0.95 \\ S_0 + \dfrac{1 - S_0}{2} & S_0 \geq 0.95 \end{cases};$$

S_0 – начальная емкость, которая выбрана с помощью модели ансамбля деревьев;

m – медиана значения для тестовой выборки.

Моделями xgboost удалось достаточно хорошо спрогнозировать вероятность отпада, о чем свидетельствуют значения AUC для моделей (см. таблица 1).

Таблица 1 – AUC на тренировочной и тестовой избирателях

Тип учреждения	Тренировочная выборка	Тестовая выборка
Банк	86,6%	86,5%
МФО	82%	77%

Сравнение с моделями логистической регрессии (М2-один уровень лаговости, М3-2 уровня лаговости переменных). Как видим из таблицы 2 даже увеличение лагов в регресорах логистических моделей не приводит к значительному покрещення их точности. Модели бустингових ансамблей деревьев значительно лучше справляются с задачей.

Таблица 2 – Сравнение с моделями логистической регрессии

	M1	M2	M3
Банк	86%	71%	75%
МФО	82%	70%	72%

Заключение

В целом данный подход к моделированию позволяет достигнуть высокой точности оценки вероятности отпада с помощью тюнига бустинговых ансамблей деревьев. Чувствительность этих моделей к такой настройке, а также ярко выраженные оптимальные значения параметров позволяют достаточно гибко строить модель под любые нужды и данные.

Библиография

1. Kim, Ae Chan, Seongkon Kim, Won Hyung Park, and Dong Hoon Lee. 2014. “Fraud and Financial Crime Detection Model Using Malware Forensics.” Multimedia Tools and Applications 68(2): 479–96. https://doi.org/10.1007/s11042-013-1410-3.
2. Carminati, Michele et al. 2014. “BankSealer: An Online Banking Fraud Analysis and Decision Support System.” In ICT Systems Security and Privacy Protection, eds. Nora Cuppens-Boulahia et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 380–94.
3. Hand, D J et al. 2008. “Performance Criteria for Plastic Card Fraud Detection Tools.” Journal of the Operational Research Society 59(7): 956–62. https://doi.org/10.1057/palgrave.jors.2602418.
4. Molloy, Ian et al. 2017. “Graph Analytics for Real-Time Scoring of Cross-Channel Transactional Fraud.” In Financial Cryptography and Data Security, eds. Jens Grossklags and Bart Preneel. Berlin, Heidelberg: Springer Berlin Heidelberg, 22–40.
5. Adewumi, Aderemi O, and Andronicus A Akinyelu. 2017. “A Survey of Machine-Learning and Nature-Inspired Based Credit Card Fraud Detection Techniques.” International Journal of System Assurance Engineering and Management 8(2): 937–53. https://doi.org/10.1007/s13198-016-0551-y.
6. Jog, Anita, and Anjali A Chandavale. 2018. “Implementation of Credit Card Fraud Detection System with Concept Drifts Adaptation.” In Intelligent Computing and Information and Communication, eds. Subhash Bhalla et al. Singapore: Springer Singapore, 467–77.

7. Khattri, Vipin, and Deepak Kumar Singh. 2018. “A Novel Distance Authentication Mechanism to Prevent the Online Transaction Fraud.” In Advances in Fire and Process Safety, eds. N A Siddiqui, S M Tauseef, S A Abbasi, and Ali S Rangwala. Singapore: Springer Singapore, 157–69.
8. Blackwell, Clive. 2014. “Using Fraud Trees to Analyze Internet Credit Card Fraud.” In Advances in Digital Forensics X, eds. Gilbert Peterson and Sujeet Shenoi. Berlin, Heidelberg: Springer Berlin Heidelberg, 17–29.

The estimation of the probability of outflow of client

Dmitrii V. Lyutyagin

PhD in Economics,
Associate Professor,
Russian State Geological Prospecting University,
117485, 23, Miklukho-Maklaya st., Moscow, Russian Federation;
e-mail: l-d-v@list.ru

Yurii V. Zabaikin

PhD in Economics,
Associate Professor,
Russian State Geological Prospecting University,
117485, 23, Miklukho-Maklaya st., Moscow, Russian Federation;
e-mail: 89264154444@yandex.ru

Abstract

Construction of the model began with censoring and generation of independent variables and target field. Censoring occurred by forming a target field based on whether the client will return after a certain fixed day of operation.

Next, the target field was built according to the following scheme:

- if the client performed an operation on the day of observation, this observation is taken into account in the sample for the contribution;

- if the client made an operation on the day of observation, this observation is taken into account in the sample for the Bank;

Thus, one observation can fall into both samples.

The next logical step is to form an "absence measure" as a break in days between two consecutive observations. In this case, a pass is considered significant if its absence measure satisfies 3 conditions:

- if the client was absent for more than 3 days;

- if the value of the absence measure is extreme, namely falls into the 0.975 quartile of the distribution of all the absence measures that were observed in a particular client;

- the maximum break between operations is at least 15 days.

To build the models, a number of lag variables were generated, that is, the value of observations with a lag of one day. If the client was absent on that day, a linear approximation was made between the two closest observations of his activity.

For citation

Lyutyagin D.V., Zabaikin Yu.V. (2019) Veroyatnost' ottoka klienta pri realizatsii skoringovoi modeli v usloviyakh deyatel'nosti prirodokhozyaistvennogo predpriyatiya [The estimation of the probability of outflow of client]. *Ekonomika: vchera, segodnya, zavtra* [Economics: Yesterday, Today and Tomorrow], 9 (5B), pp. 543-550.

Keywords

Model, client, variables, logical step, operation.

References

1. Kim, Ae Chan, Seongkon Kim, Won Hyung Park, and Dong Hoon Lee. 2014. "Fraud and Financial Crime Detection Model Using Malware Forensics." Multimedia Tools and Applications 68 (2): 479–96. https://doi.org/10.1007/s11042-013-1410-3.
2. Carminati, Michele et al. 2014. "BankSealer: An Online Banking Fraud Analysis and Decision Support System." In ICT Systems Security and Privacy Protection, eds. Nora Cuppens-Boulahia et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 380–94.
3. Hand, D J et al. 2008. "Performance Criteria for Plastic Card Fraud Detection Tools." Journal of the Operational Research Society 59 (7): 956–62. https://doi.org/10.1057/palgrave.jors.2602418.
4. Molloy, Ian et al. 2017. "Graph Analytics for Real-Time Scoring of Cross-Channel Transactional Fraud." In Financial Cryptography and Data Security, eds. Jens Grossklags and Bart Preneel. Berlin, Heidelberg: Springer Berlin Heidelberg, 22–40.
5. Adewumi, Aderemi O, and Andronicus A Akinyelu. 2017. "A Survey of Machine-Learning and Nature-Inspired Based Credit Card Fraud Detection Techniques." International Journal of System Assurance Engineering and Management 8 (2): 937–53. https://doi.org/10.1007/s13198-016-0551-y.
6. Jog, Anita, and Anjali A Chandavale. 2018. "Implementation of Credit Card Fraud Detection System with Concept Drifts Adaptation." In Intelligent Computing and Information and Communication, eds. Subhash Bhalla et al. Singapore: Springer Singapore, 467–77.
7. Khattri, Vipin, and Deepak Kumar Singh. 2018. "A Novel Distance Authentication Mechanism to Prevent the Online Transaction Fraud." In Advances in Fire and Process Safety, eds. N A Siddiqui, S M Tauseef, S A Abbasi, and Ali S Rangwala. Singapore: Springer Singapore, 157–69.
8. Blackwell, Clive. 2014. "Using Fraud Trees to Analyze Internet Credit Card Fraud." In Advances in Digital Forensics X, eds. Gilbert Peterson and Sujeet Shenoi. Berlin, Heidelberg: Springer Berlin Heidelberg, 17–29.