ОТБОР ПЕРЕМЕННЫХ ДЛЯ АНАЛИЗА И ПРОГНОЗИРОВАНИЯ НЕСТАБИЛЬНОСТИ С ПОМОЩЬЮ МОДЕЛЕЙ ГРАДИЕНТНОГО БУСТИНГА*

Сергей Георгиевич Шульгин
Российская академия народного хозяйства и государственной службы при Президенте РФ;
Национальный исследовательский университет «Высшая школа экономики»

В статье предлагается метод для отбора переменных, наиболее важных для анализа и прогнозирования социально-политической нестабильности. В качестве источника данных по нестабильности используются данные The Cross-National Time Series (CNTS), Global Terrorism Database (GTB) *и базы данных государственных переворотов. В качестве источников данных для независимых переменных мы используем данные* World Bank, United Nation Population Division, Polity IV, Maddison Database, Worldwide Governance Indicators *и др. Анализируемое множество факторов содержит около 250 показателей. Для отбора переменных мы используем метод градиентного бустинга, с помощью которого мы ранжируем переменные по их важности для анализа и предсказания различных измерений социально-политической нестабильности. Среди наиболее важных переменных выделяются переменные, которые описывают историю существования и устойчивости режима («Долговечность режима», «Возраст государственности и независимости»), переменные характеризующие тип режима («Индекс регулирования участия в политике», «Комбинированная оценка* Polity IV*»), переменные, характеризующие структуру населения и занятости («Население в возрасте от 0 до 4 лет», «Доля занятых в промышленности»), переменные отражающие состояние мировой конъюнктуры («Цена на золото», «Индекс потребительских цен»).*

Существуют различные подходы к анализу факторов, влияющих на социально-политическую нестабильность. Традиционным подхо-

* Исследование выполнено при поддержке Российского научного фонда (проект № 18-18-00254).

дом является экспертный отбор факторов, которые значимо влияют на различные виды нестабильности.

Среди факторов, которые могут оказывать влияние на дестабилизационные социально-политические процессы, можно выделить несколько групп факторов.

1. Социально-экономические характеристики страны. Среди них можно выделить макроэкономические показатели, такие как ВВП, ВВП на душу населения и их темпы роста, дефицит бюджета, показатели неравенства, инфляция, уровень сбережений и инвестиционной активности, показатели внешней торговли и множество других.

2. Социокультурные характеристики общества – такие как уровень образования, характеристики национального состава, характеристики, описывающие распространение и роль различных религий, уровень доверия в обществе, приверженность тем или иным ценностям, роль и активность отдельных классов, уровень распространения технологий, характеристики, описывающие распространенность различных языков и многие другие.

3. Характеристики, описывающие модель социального и политического устройства: такие как тип политического режима, тип правовой системы, характеристики формальных и неформальных институтов, характеристики системы государственного управления.

4. Демографические факторы – факторы, связанные с численностью населения и ее структурным составом, с долей в общей численности отдельных групп населения, например, с долей молодежи, интенсивностью миграционных процессов (как внутренних, так и внешних), интенсивностью процессов рождаемости и смертности.

5. Географические и исторические факторы: описывающие положение, взаимное расположение, совместное историческое развитие.

Любая классификация не будет полной или достаточно строгой, так как существуют факторы, которые будут относиться сразу ко многим типам и влиять через разные механизмы: например, распространение информационных технологий, одновременно влияет и на макроэкономические показатели, и на социально-экономические характеристики общества и т. п.

В зависимости от того, как авторы подходят к анализу процессов социально-политической нестабильности, они подчеркивают

важность тех или иных факторов (Esty, Goldstone, Gurr *et al.* 1998; Цирель 2012, 2015; Коротаев, Зинькина 2011, 2012; Малков и др. 2012; Коротаев и др. 2017; Васькин и др. 2018; Korotayev *et al.* 2011; 2018; Slinko *et al.* 2017; Przeworski *et al.* 2000). Обзор некоторых подходов к анализу нестабильности можно найти в нашей работе 2017 г. (Korotayev, Shulgin, Zinkina 2017).

Другим подходом к анализу социально-политической нестабильности является применение методов машинного обучения и анализа больших данных. Существует подход, при котором для анализа процессов нестабильности используются данные, собранные на микроуровне (отдельные события высокой географической детализации, поведение отдельных людей и т. п.); для работы с ними используются модели, условно называемые методами машинного обучения. Обзор подходов анализу представлен в работе 2017 г. (Donnay 2017), а в качестве примеров подобных исследований могут служить такие работы, как Connelly *et al.* 2016; Donnay *et al.* 2016; Coppock *et al.* 2016 и ряд других.

В настоящей работе анализируются данные собранные на страновом уровне, что позволяет работать с широким набором независимых факторов, собранных из разных источников.

Методология

Мы выделяем в отдельную группу зависимых переменных набор факторов, которые отражают те или иные формы социально-политической нестабильности. Для поиска наиболее важных факторов мы тренируем (оцениваем) множество моделей, в каждой из которых лишь одна из зависимых переменных используются в качестве целевой объясняемой переменной. При этом ни одна из зависимых переменных никогда не включается в число факторов в оцениваемые модели.

Модель

Для заданного набора данных D определены n точек данных, в котором каждая точка данных – это набор из объясняемой (зависимой) переменной y_i и множества из m независимых факторов X_i:

$$D = \{(y_i, X_i)\} \ (|D| = n, X_i \in \mathbb{R}^m, y_i \in \mathbb{R}), \quad (1)$$

где $\mathbb{R}$ – стандартное обозначение для множества действительных чисел.

В такой формулировке, наша задача среди всего множества независимых факторов X, выделить необходимое подмножество, то есть отдельные его факторы, которые оказываются наиболее важными для объяснения y.

В данной работе мы используем метод, при котором мы пытаемся найти оценку зависимой переменной y_i в форме K аддитивных функций

$$\hat{y}_i = \sum_{k=1}^{K} f_k(X_i) \quad (2)$$

$f_k(X_i)$ – функция, которая принадлежит к подмножеству классификационных и регрессионных деревьев (CART – Classification and Regression Tree).

Класс функций, которые определяются как:

$$CART = \{f(X) = w_{q(X)}\} \ (q: \mathbb{R}^m \rightarrow T; w \in \mathbb{R}^T), \quad (3)$$

где q(X) – описывает дерево, вершинами которого являются правила относительно значений X. Функция q(X) ставит в соответствие для определенной точке данных X_i, определенный лист (конечную вершину) (T). Листья в CART описывают результат классификации, которым присвоены веса w. Аппроксимирующая функция $f_k(X)$ определяется структурой дерева q(X) и весами листьев w.

Процесс обучения (тренировки) модели сводится к минимизации функционала L, в которой суммируется ошибка между оцененными ($\hat{y}_i$) и реальными значениями (y_i) зависимой переменной, а также учитывается сложность (размерность) CART функции. Вторая часть функционала L – это элемент так называемой регуляризации, подход с помощью которого мы контролируем сложность CART функции и пытаемся найти самую простую структуру из возможных CART функций.

Для минимизации функционала L используется последовательный (итеративный) процесс, где на каждой итерации оценивается градиент в направлении минимизации L. Подробнее описание функционала и алгоритма оптимизации см.: Chen, Guestrin 2016.

Использование моделей градиентного бустинга (GBM) не требует нормализации данных для корректной работы и хорошо работает без предварительной обработки входных данных. В работе мы использовали GBM для всех оценок, которые производили с помощью библиотеки XGBoost (Chen, Guestrin 2016).

Данный метод успешно применяется для широкого класса задач, связанного с отбором наиболее важных переменных в задачах с высокой размерностью. Например, в отборе оптимальных характеристик соискателей для подбора им наиболее интересных и релевантных вакансий (Volkovs, Wei Yu, Poutanen 2017), или предсказания о том какие наиболее значимые аффилиации авторов влияют на факт, что их статьи принимаются на основные авторитетные научные конференции в области машинного обучения, больших данных и т. п. (Sandulescu, Chiru 2016), или анализе физических данных CERN, полученных на большом адронном коллайдере, в попытках найти факторы, влияющие на вероятность наблюдения редкого физического явления распада тау-лептона на три мюона ($\tau \rightarrow 3\mu$) (Mironov, Guschin 2015) и во многих других приложениях.

Данные

В качестве исходных данных о нестабильности мы используем данные *Cross National Time Series* (*CNTS*), *Global Terrorism Database* (*GTB*) и базы данных государственных переворотов.

База данных The Cross National Time Series (CNTS) – это результат работы по сбору и систематизации данных, начатой Артуром Банксом (Banks, Wilson 2018) в 1968 г. в Университет штата Нью-Йорк в Бингемтоне через обобщение архива данных *The Statesman's Yearbook*, публикуемых с 1864 г. В базе содержаться данные по более чем 200 странам. База данных содержит годовые значения переменных, начиная с 1815 г. В базе данных исключены периоды двух мировых войн 1914–1918 и 1940–1945 гг.

В данной работе мы используем в качестве зависимых переменных данные, описывающие различные аспекты внутренних конфликтов (*domestic*). Эти данные получены из анализа страновых событий по 8 различным подкатегориям:

- Политические убийства (Assassinations, domestic1).
- Политические забастовки (General Strikes, domestic2).
- Партизанские действия (Guerrilla Warfare, domestic3).
- Правительственные кризисы (Government Crises, domestic4).

- Политические репрессии (Purges, domestic5).
- Массовые беспорядки (Riots, domestic6).
- Перевороты и попытки переворотов (Revolutions, domestic7).
- Антиправительственные демонстрации (Anti-Government Demonstrations, domestic8).

К Политическим убийствам (Assassinations, domestic1) относятся любые политически мотивированным убийства или покушения на убийства высших правительственных чиновников или политиков.

К Политическим забастовкам (General Strikes, domestic2) относятся забастовки, в которых участвовало 1000 или более работников, более одного работодателя и при этом были требования, направленные против национальной политики, правительства или органов власти.

К Партизанским действиям (Guerrilla Warfare, domestic3) относится любая вооруженная деятельность, диверсии или взрывы, совершаемых независимыми группами граждан или нерегулярными вооруженными силами, которые направлены на свержение нынешнего режима.

К Правительственным кризисам (Government Crises, domestic4) относятся любые ситуации, которые грозят привести к падению текущего режима – за исключением вооруженных переворотов, напрямую направленные на это.

К Политическим репрессиям (Purges, domestic5) относятся любые систематические устранения политической оппозиции (лишения свободы или убийства) среди действующих членов режима или политической оппозиции.

К Массовым беспорядкам (Riots, domestic6) относятся любые протесты или столкновения, связанные с использованием насилия, в которых принимали участие более 100 граждан.

К Переворотам и попыткам переворотов (Revolutions, domestic7) относятся любые незаконные или связанные с принуждением изменения в правящей элите, а также любые попытки таких изменений. Переменная «Перевороты и попытки переворотов» также учитывает все удачные и неудачные вооруженные восстания, целью которых является получение независимости от центрального правительства.

К Антиправительственным демонстрациям (Anti-Government Demonstrations, domestic8) относятся любые мирные публичные собрания, в которых принимает участие 100 человек и более, а в качестве основной цели проведения является выражение несогласия с

политикой правительства или власти за исключением демонстраций с выраженной направленностью против иностранных государств.

Все перечисленные 8 подкатегорий используются при построении общего индекса социально-политической стабилизации (domestic9). Для этого составители базы данных CNTS присвоили каждой подкатегории определенный вес (см. Табл. 1).

Табл. 1. Веса подкатегорий, используемых при построении индекса социально-политической стабилизации

Подкатегория	Название переменной	Вес в индексе социально-политической стабилизации (domestic9)
Политические убийства (Assassinations)	cnts_domestic1	25
Политические забастовки (General Strikes)	cnts_domestic2	20
Партизанские действия (Guerrilla Warfare)	cnts_domestic3	100
Правительственные кризисы (Government Crises)	cnts_domestic4	20
Политические репрессии (Purges)	cnts_domestic5	20
Массовые беспорядки (Riots)	cnts_domestic6	25
Перевороты и попытки переворотов (Revolutions)	cnts_domestic7	150
Антиправительственные демонстрации (Anti-Government Demonstrations)	cnts_domestic8	10

Индекс социально-политической стабилизации (Weighted Conflict Measure, domestic9) рассчитывается по формуле (4):

$$domestic9 = \frac{\sum_{i=1}^{8} w_i \, \text{cnts\_}domestic_i}{8} * 100, \quad (4)$$

где w_i – веса, приведенные в последнем столбце Табл. 1.

Кроме показателя domestic9 для анализа мы построили переменную domestic9 с лагом (cnts_domestic9_prev), которая показывает общее значение страновой нестабильности в предыдущем году. Также мы построили упреждающую переменную (cnts_domestic9_next) для оценки общего уровня нестабильности в будущем году.

Помимо данных CNTS, в качестве объясняемой переменной мы используем два индикатора из Global Terrorism Database (START 2016). Мы используем переменные:

n_terror_attack – количество террористических атак,

Nkill – количество убитых.

База содержит данные с 1970 (в анализируемой версии по 2015 г. включительно).

Из базы данных государственных переворотов (Marshall 2016) для независимых переменных мы взяли для анализа переменную: coup_detat_failed_coup_detat – государственные перевороты и попытки переворотов (аналог переменной cnts_domestic8).

База данных государственных переворотов охватывает временной период с 1960 по 2016 г.

Всего в качестве зависимых (объясняемых, целевых) для данного анализа было отобрано 14 переменных. Все зависимые переменные мы представили в форме бинарного классификатора, с помощью которого моделировалось наличие или отсутствие в данном году, в данной стране нестабильности по анализируемой переменной. Точки данных, в которых значение переменной было больше 0 были классифицированы как факт нестабильности. Для переменной n_terror_attack пороговым значением было выбрано N = 50.

В Табл. 2 приведена статистика по 14 зависимым переменным и число случаев нестабильности и отсутствия нестабильности.

Табл. 2. Статистика по числу случаев нестабильности (ее отсутствия)

Переменная	Общее число N	Число случаев нестабильности	Число случаев отсутствия нестабильности	Пропущенные данные
cnts_domestic1	12 198	1046	11 152	3924
cnts_domestic2	12 198	936	11 262	3924
cnts_domestic3	12 198	1611	10 587	3924
cnts_domestic4	12 198	1618	10 580	3924
cnts_domestic5	12 198	1124	11 074	3924

Окончание табл. 2

Переменная	Общее число N	Число случаев нестабильности	Число случаев отсутствия нестабильности	Пропущенные данные
cnts_domestic6	12 198	2296	9902	3924
cnts_domestic7	12 198	1618	10 580	3924
cnts_domestic8	12 198	2517	9681	3924

Переменная	Общее число N	Число случаев нестабильности	Число случаев отсутствия нестабильности	Пропущенные данные
cnts_domestic9	12 198	5663	6535	3924
cnts_domestic9_next	10 789	4839	5950	5333
cnts_domestic9_prev	10 736	4814	5922	5386
coup_detat_failed_coup_detat	16 094	465	15 629	28
n_terror_attack	3491	527	2964	12 631
nkill	3447	2368	1079	12 675

Выбор параметров модели и тренировка моделей

Для оценки (тренировки) модели градиентного бустинга необходимо выбрать набор параметров, определяющих работу алгоритма. Одной из главных проблем, которые необходимо решить при оценке – это проблема переобучения модели (over-fitting). Переобучение выражается в том, что при большом количестве данных и степеней свободы модель может очень точно описать существующие закономерности на обучающей выборке (training set), однако полученные закономерности могут оказаться неприменимы за пределами обучающей выборки.

Для решения этой проблемы мы используем подход кросс-валидации (cross-validation), когда из имеющихся данных выделяем обучающую (train) и тестовую (test) выборки. Обучающая выборка используется для тренировки моделей. Тестовая выборка используется только для анализа полученных результатов (и не участвует в процессе обучения). Процесс обучения модели, итеративный и на каждой итерации мы анализируем качество полученной оценки на тестовой выборке и принимаем решение об остановке дальнейшего обучения модели в случае, когда за определенное число последних итераций не произошло улучшение результатов оценки в тестовой выборке.

Тестовую и обучающую выборки мы формируем с помощью процедуры k-fold кросс-валидации, когда вся выборка разбивается на k случайных частей (Kuhn 2008) и одна из этих используется в качестве тестовой, а остальные k-1 в качестве обучающей. Процедура оценки модели с этим разбиением повторяется k раз, так чтобы каждая из k-частей побывала тесовой выборкой.

В данной работе мы выбрали k равным 5 и для каждой зависимой переменной провели оценку модели с помощью процедуры кросс-валидации 20 раз (каждый раз с новым разбиением на 5 случайных подвыборок). В результате для каждой из 14 зависимых переменных мы получили 100 оценок моделей, в каждой из которых оценивается значимость независимых факторов.

Складывая для каждого независимого фактора оценки его важности во всех 100 оценках модели, мы получаем результирующую агрегированную оценку значимости каждого независимого фактора для анализируемой зависимой переменной.

Параметр глубины деревьев (*max.depth*) эмпирическим путем выбран равным 5. Уменьшение упрощает структуру (желаемое свойство), однако качество оценки модели падает. Увеличение глубины в целом улучшает качество оценки модели – но только для обучаемой выборки, в тестовой выборке улучшений качества оценки с увеличением глубины не наблюдается. Полученные результаты робастны относительно большого диапазона параметров глубины.

Параметр скорости сходимости (*eta*) был выбран 0,15. Анализировался различный диапазон параметров скорости сходимости. Более низкие значения потенциально позволяют достигнуть более высокой точности модели. С учетом контроля процесса обучения на тестовой выборке, этот параметр влияет, в основном, на скорость работы. В результате более низкие значения *eta* не дают выигрыша в точности модели в тестовой выборке, и замедляют процесс тренировки.

Использовалась функция ошибок, в которой оценивается доля полученных оценок, отличающихся от истинных (наблюдаемых) значений.

Как уже упоминалось выше, каждый процесс обучения контролировался на тестовой выборке. Для параметра *early_stopping_ rounds* мы использовали значение 50.

Тренировка модели – это итеративная процедура построения классификатора, когда на каждом шаге к существующему классу классификаторов добавляется новая CART функция, так чтобы минимизировать ошибку оценки зависимой переменной (стараясь поддерживать максимально простую структуру CART).

На Рис. 1 представлена динамика функции ошибок. По горизонтальной оси откладывается номер итерации, по вертикальной оси – ошибки классификации на данной итерации.

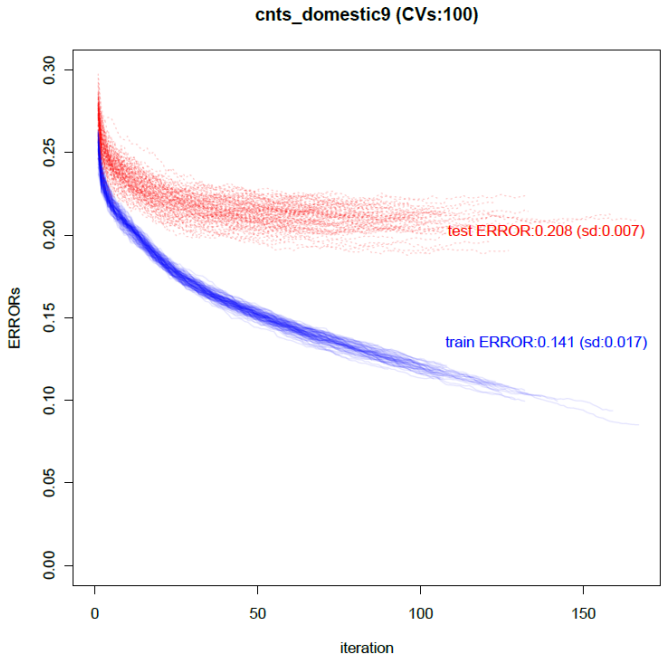


Рис. 1. Динамика ошибок обучающей и тестовой выборки для Индекса социально-политической стабилизации (cnts_domestic9)

Примечание: на графике приводится среднее значение и стандартное отклонение (в скобках) для ошибок рассчитанное для 100 оценок.

На Рис. 1 приводятся результаты оценок для 100 различных кросс-валидаций – каждой из которой соответствует своя кривая ошибок обучающей (синяя сплошная линия) и тестовой выборки (красная прерывистая линия). На рисунке 1 также приводится среднее значение, полученное по всем 100 оценкам. Так для обучающей выборки (*train Error*) – это значение равно 0,141. То есть на обучающей выборке модель в среднем, дает оценку индекса социально-политической стабилизации, которая ошибается 14,1 % случаев (относительно реального значения индекса социально-политической стабилизации). Можно заметить, что с ростом числа итерации (усложнения модели) ошибка на обучающей выборке (синие сплошные кривые) стабильно снижается и если продолжать обучение можно достигнуть более высокой точности, однако для нас

критерием точности модели является ошибка на тестовой выборке (*test-error*) – в среднем для переменной cnts_domestic9 она оказывается равной 0.208. С использованием параметра *early_stopping_rounds* мы прерываем дальнейшую тренировку модели, если за последние 50 итераций не было улучшений в оценках на тестовой выборке (*test-error*).

Качество полученных оценок, мы также можем оценить с помощью ROC-кривых, которые показывают, как соотносятся частота ложно-положительных оценок (*false-positive rate – FPR*) с частотой истинно-положительными оценок (*true-positive rate – TPR*). На Рис. 2 представлены ROC-кривые для 100 оценок модели на обучающих и тестовых выборках для Индекса социально-политической стабилизации (*cnts_domestic9*).

На Рис. 2 по горизонтальной оси отложен коэффициент специфичности (*Specificity*), который рассчитывается как единица минус частота ложно-положительных оценок (1 – FPR), а по вертикальной оси значение коэффициента чувствительности (S*ensitivity*) который равен коэффициенту истинно-положительных оценок (TPR). Идеальный классификатор – это классификатор, у которого сочетаются нулевые ложно-положительные ошибки (специфичности = 1) и 100 % чувствительность (то есть TPR = 1). Модель, которая случайным образом пытается угадать значение, будет соответствовать горизонтальной прямой *Sensitivity* = 1 – *Specificity* (или TPR = FPR). ROC кривая для модели описывает сочетания TPR и FPR для полученных оценок. Чем ближе ROC кривая к горизонтальной прямой, тем хуже работает модель – тем меньше объясняющая способность модели и ниже качество полученных оценок. В качестве интегрального критерия качества модели на ROC-кривой используется показатель *площади под ROC-кривой* (*AUC* – Area Under Curve). Для идеального классификатора AUC = 1, для абсолютно не информативной модели (горизонтальной прямой) AUC = 0,5. На рисунке 2 видно, что модель на обучающей выборке обладает высокими прогностическими возможностями, но и на тестовой выборке ее прогностические способности оказываются не нулевые (AUC = 0,856).

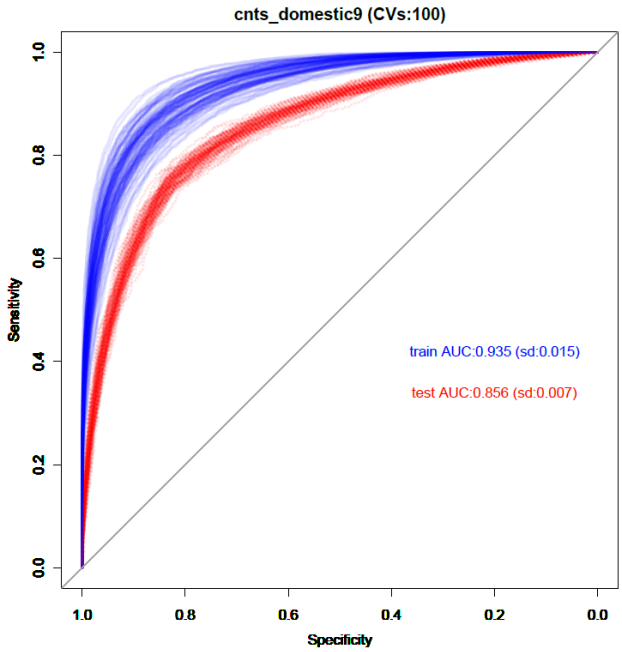


Рис. 2. ROC-кривые для оценок модели на обучающих и тестовых выборках для Индекса социально-политической стабилизации (cnts_domestic9)

Примечание: на графике приводится среднее значение и стандартное отклонение (в скобках) показателя AUC рассчитанное для 100 прогнозных оценок.

Полученные результаты оценок ошибок и AUC для выбранных нами 14 измерений нестабильности приведены в Табл. 3, а подробные графики динамики ошибок и ROC кривые для всех 14 зависимых переменных, приводятся в приложении 1 и 2 соответственно.

Табл. 3. Оценки качества моделей на обучающих и тестовых выборках

Переменная	Обучающая выборка (train)		Тестовая выборка (test)	
	Error	AUC	Error	AUC
cnts_domestic1	0,065 (0,009)	0,934 (0,036)	0,079 (0,005)	0,814 (0,019)
cnts_domestic2	0,058	0,930	0,072	0,834

Переменная	Обучающая выборка (train)		Тестовая выборка (test)	
	Error	AUC	Error	AUC
	(0,010)	(0,052)	(0,005)	(0,033)
cnts_domestic3	0,051 (0,015)	0,979 (0,014)	0,093 (0,005)	0,904 (0,012)
cnts_domestic4	0,094 (0,012)	0,920 (0,037)	0,121 (0,007)	0,805 (0,018)
cnts_domestic5	0,071 (0,007)	0,927 (0,033)	0,083 (0,005)	0,871 (0,022)
cnts_domestic6	0,107 (0,020)	0,936 (0,031)	0,151 (0,007)	0,840 (0,012)
cnts_domestic7	0,061 (0,012)	0,972 (0,013)	0,100 (0,006)	0,884 (0,010)
cnts_domestic8	0,109 (0,016)	0,935 (0,022)	0,157 (0,007)	0,837 (0,010)
cnts_domestic9	0,141 (0,017)	0,935 (0,015)	0,208 (0,007)	0,856 (0,007)
cnts_domestic9_next	0,149 (0,022)	0,930 (0,018)	0,222 (0,009)	0,841 (0,008)
cnts_domestic9_prev	0,141 (0,020)	0,936 (0,018)	0,217 (0,009)	0,847 (0,009)
coup_detat_failed_coup_detat	0,021 (0,002)	0,940 (0,026)	0,026 (0,003)	0,896 (0,019)
n_terror_attack	0,009 (0,009)	0,999 (0,002)	0,059 (0,009)	0,955 (0,014)
nkill	0,138 (0,035)	0,920 (0,044)	0,236 (0,014)	0,796 (0,024)

Примечание: приводится среднее значение и стандартное отклонение (в скобках) ошибок и показателя AUC рассчитанное для 100 прогнозных оценок.

Оценка значимости факторов

После тренировки (оценки) моделей градиентного бустинга (по 100 моделей на каждую из 14 зависимых переменных), для каждой полученной оценки мы анализировали значимость независимых факторов. Важность факторов в модели градиентного бустинга оценивается по 3 параметрам:

- *gain* – описывает относительный вклад соответствующего фактора в модель, рассчитанный путем оценки вклада фактора для каждого дерева в модели;

• *cover* – оценивает, для какой доли в исходных данных (для скольких точек данных) анализируемая переменная влияет на результат классификации;

• *frequency* – оценивает сколько раз независимый фактор используется для разделения данных по всем деревьям.

Каждый из этих факторов может быть использован для оценки относительной важности переменной, однако параметр *frequency* является упрощенной версией *gain* (без учета вклада переменной в результат классификации, поэтому мы его напрямую не используем при учете значимости, так как эта информации уже заложена в показателе *gain*). Итоговый индекс значимости переменных представлен в формуле (5).

$$importance = a*gain + (1-a)*cover,\ a \in [0, 1] \quad (5)$$

Коэффициент *a* перед переменной *gain* показывает, насколько важным будет вклад переменной в объясняющую способность модели, (1-a) – показывает, насколько мы считаем, что покрытие данными отражает значимость факторов.

В данной работе мы принимает $a = 2/3$ – это означает, что вклад в объясняющую способность (*gain*) мы оцениваем в 2 раза выше, чем покрытие данными (*cover*). Переменные *cover и gain,* сильно коррелированы (с корреляцией больше 95 %, рассчитанной между всеми 200+ переменными после агрегирования *gain* и *cover* по 100 кросс-валидациям).

Размерность коэффициента важности такова, что количественно коэффициент важности (*importance*) на уровне 1 можно интерпретировать как 1 % вклад в объясняющую способность модели (с учетом вклада в покрытие).

Описание результатов

Результаты анализа важности независимых факторов мы представляем следующим образом. Сначала мы представляем агрегированные результаты важности всех факторов для всех 14 зависимых переменных. Далее кратко перечисляем наиболее важные факторы по каждому из 14 направлений дестабилизации.

По результатам оценки по 100 моделей для каждой из 14 переменных мы выделили следующие наиболее важные независимые факторы:

1. Население (cnts_pop1 [8.6]).
2. Долговечность режима (p_durable [4.1]).

3. Индекс регулирования участия в политике (p_parreg [2.3]).

4. Плотность населения (cnts_pop_density [1.9]).

5. Доля занятых в промышленности (cnts_percent_work_force_in_industry [1.6]).

6. Возраст государственности, независимость (nationality_sovereignty [1.5]).

7. Комбинированная оценка Polity IV (p_polity_2_2 [1.5]).

8. Индекс потребительских цен (consumer_price_index [1.4]).

9. Население в возрасте от 0 до 4 лет (за 5 лет) (X0__4 [1.3]).

10. Цена на золото (Gold_price [1.3]).

В квадратных скобках приводится показатель важности (importance) данной переменной по итогам оценки всего множества моделей (для всех измерений дестабилизации).

Вторая по важности независимая переменная «Долговечность режима» (p_durable [4.1]), это переменная из базы данных полити-4 (Polity IV 2015) оценивает число лет с момента последнего изменения режима (определяемое трехуровневым изменением показателя p_polity). Переменная измеряется в годах и оценивает, как долго не менялся тип режима.

Третья по важности переменная «Индекс регулирования участия в политике» (p_parreg [2.3]), также переменная из базы данных полити-4 (Polity IV 2015). Эта переменная оценивает ограничения на выражения политических предпочтений. Переменная измеряется 5 уровнями от значений «не регулируемое» до «максимально зарегулированные» (подробнее см. описание базы данных полити-4).

Пятая независимая переменная «Возраст государственности, независимость» (nationality_sovereignty [1.5]), была построена нами самостоятельно и оценивает, сколько прошло лет с момента обретения государством национальной независимости.

Седьмая независимая переменная «Комбинированная оценка Polity IV» (оценка типа режима из базы данных Fridom House (Freedom in the World 2015).

Для большинства моделей общая численность населения является исключительно важным фактор нестабильности – это связано с тем, что в основном мы используем показатели нестабильности в абсолютном выражении на основе событийного подхода и в более крупных странах больше шансов, что будут зарегистрированы (замечены) события, относящиеся к нестабильности. Важным фактором также является плотность населения, что говорит о том, что

нестабильность связано не только с абсолютной численностью населения.

Среди наиболее важных переменных можно выделить переменные которые описывают историю устойчивости режима (Долговечность режима, Возраст государственности[1]), несколько переменных характеризующих тип режима (Индекс регулирования участия в политике, Комбинированная оценка Polity IV), несколько переменных описывающих структуру населения и структуру занятости (Население в возрасте от 0 до 4 лет (за 5 лет), Доля занятых в промышленности), переменные отражающие мировую конъюнктуру (Цена на золото, Индекс потребительских цен).

Также мы выделили следующие по важности 15 факторов: Доля валовых внутренних инвестиций в ВВП (gross_fixed_capital_formation_percent_gdp), Количество городов на душу населения (cnts_population_cities_over_per_capita), Экспорт на душу населения, в долларах (cnts_exports_per_capita), Эффективность законодательной власти (cnts_effectiveness_of_legislature), Темпы роста ВВП по ППС (gdp_PPP_annual_growth), Население (population), Индекс фракционности (cnts_polit01), Цена на нефть марки Brent в номинальных ценах (Brent_price_nominal), ВВП, в текущей национальной валюте (gdp_current_LCU), Темпы роста ВВП на душу населения по ППС, в постоянных ценах 2011 г. (growth_rate_gdp_per_capita_PPP_WB), Темпы роста ПИИ, приток (fdi_inward_percent_gdp), Доля грамотных (cnts_percent_literate), Площадь (cnts_area1), Возраст государственности, появление государства или квазигосударственного образования (nationality_appearance), ВВП на душу населения по ППС, в постоянных ценах 2011 г. (gdp_per_capita_PPP_Mad).

Все перечисленные факторы учитывают 42 % общей важности всех факторов, группа из вторых по важности 15 факторов, объясняет 17 % от общей важности всех факторов, которые можно интерпретировать как вклад в объясняющую способность во всех оцененных моделях.

Для отдельных направлений дестабилизации мы приводим важность факторов с разбиением на три группы (по степени важности):

---

[1] Отметим, что в одной из статей в предыдущем выпуске Мониторинга уже было показано, что небольшой возраст государственности является достаточно мощным статистически значимым предиктором социально-политической дестабилизации (Гринин и др. 2017).

В первой группе 5 наиболее важных переменные. Для этих переменных в квадратных скобках, после кодового названия переменной, приводится значение коэффициента важности (***importance)*.** Среди этих переменных значимы различия в ***importance и значим*** ранг важности.

Во второй по важности группе – переменные, для которых коэффициент ***importance*** > 1,5. Различия в параметре важности (***importance***) между переменными второй группы могут быть не значимы и ранг их важности может изменяться (если оценить их на альтернативном наборе оценок модели с кросс-валидацией). Это также относится и к перечисленной выше важности независимых переменных, оцененной совместно для всех 14 моделей.

В третьей по важности группе переменных – переменные, для которых ***importance*** > 1 (и не превышает 1,5). Различия в важности переменных в третьей группе уже не существенны, однако в среднем они значимо менее важные, чем переменной второй группы.

Мы выделили следующие основные факторы, влияющие на переменную «**Политические убийства**» (cnts_domestic1):

1. Население (cnts_pop1 [10.4]),
2. Долговечность режима (p_durable [6]),
3. Индекс регулирования участия в политике (p_parreg [4]),
4. Население (population [2.4]),
5. Плотность населения (cnts_pop_density [2.1]).

Вторая по важности группа переменных (с коэффициентом importance > 1,5): Возраст государственности, появление государства или квазигосударственного образования (nationality_appearance), Количество городов на душу населения (cnts_population_cities_over_per_capita), Городское население в возрасте от 15 до 24 лет (urban_15__24), Уровень смертности от ран или увечий, которые люди нанесли себе сами на 100 тыс. человек, 5-летние интервалы (death_rate_injuries).

Данные переменные совместно отвечают за 48 % общей значимости модели. Вторая группа факторов отвечает за 7 %. Для отдельных индексов, как правило, не перечисляем факторы третьей группы – всего на них приходится 17 % общей объясняющей способности зависимой переменной «Политические убийства».

По результатам тренировки моделей для переменной «**Политические забастовки**» (cnts_domestic2) мы выделили следующие наиболее важные независимые факторы:

1. Население (cnts_pop1 [12.1]).

2. Возраст государственности, независимость (nationality_sovereignty [5.8]).

3. Индекс регулирования участия в политике (p_parreg [2.9]).

4. Количество городов на душу населения (cnts_population_cities_over_per_capita [2.8]).

5. Темпы роста ВВП по ППС (gdp_PPP_annual_growth [2.6]).

Во вторую по важности группу (с коэффициентом importance > > 1,5) входят следующие 8 переменных: Пересмотренная оценка Polity IV (p_polity2), Темпы роста ПИИ, приток (fdi_inward_percent_gdp), Темпы роста ВВП на душу населения по ППС, в постоянных ценах 2011 г. (growth_rate_gdp_per_capita_PPP_WB), Цена на нефть марки Brent в номинальных ценах (Brent_price_nominal), Плотность населения (cnts_pop_density), Индекс институционализированной автократии (p_autoc), Доля безработных в молодежи (по данным национальных служб) (unemployment_youth_total_NAT), Индекс потребительских цен (consumer_price_index).

Данные переменные совместно отвечают за 55 % общей значимости модели. Вторая группа факторов отвечает за 16 %, третья за 13 % общей объясняющей способности модели.

Тренировка 100 бустинговых моделей кросс-валидации позволяет выделить основные независимые факторы влияющих на переменную «**Партизанские действия**» (cnts_domestic3):

1. Население (cnts_pop1 [9.5]).

2. Долговечность режима (p_durable [3.6]).

3. Площадь (cnts_area1 [2.9]).

4. Экспорт на душу населения, в долларах (cnts_exports_per_capita [2.3]).

5. Коэффициент фертильности (fert_rate [2]).

Вторая группа факторов: Плотность населения (cnts_pop_density), Население (population), ВВП, в текущей национальной валюте (gdp_current_LCU), Возраст государственности, появление государства или квазигосударственного образования (nationality_appearance), Доля грамотных (cnts_percent_literate), Индекс потребительских цен (consumer_price_index), Индекс регулирования участия в политике (p_parreg), Индекс конкурентоспособности участия (p_parcomp), Доля промышленности в ВВП (share_industry), Доля занятых в промышленности (cnts_percent_work_force_in_industry).

Приведенные факторы совместно объясняют 50 % общей значимости модели. Вторая группа факторов объясняет 17 %, третья – 12 %.

В результате оценки моделей для переменной «**Правительственные кризисы**» (cnts_domestic4) были выделены следующие факторы:

1. Долговечность режима (p_durable [5.5]).
2. Доля занятых в промышленности (cnts_percent_work_force_in_industry [4.7]).
3. Индекс институционализированной автократии (p_autoc [3.9]).
4. Индекс фракционности (cnts_polit01 [3.5]).
5. Население (cnts_pop1 [3.1]).

Вторая по важности группа переменных (с коэффициентом importance > 1,5): Темпы роста ВВП по ППС (gdp_PPP_annual_growth), Темпы роста ПИИ, приток (fdi_inward_percent_gdp), Плотность населения (cnts_pop_density), Индекс регулирования участия в политике (p_parreg), Доля военных расходов (cnts_size_military_percent), Темпы роста ВВП на душу населения по ППС, в постоянных ценах 2011 г. (growth_rate_gdp_per_capita_PPP_WB), Цена на серебро (Silver_price), Количество солнечных пятен (sunspot_numbers).

Все эти группы совместно отвечают за 46 % общей значимости модели. Вторая группа факторов отвечает за 17 % общей важности модели, третья группа за 9 %.

По результатам тренировки моделей для зависимой переменной «**Политические репрессии**» (cnts_domestic5) мы выделили 5 наиболее важных факторов:

1. Цена на золото (Gold_price [10.7]),
2. Долговечность режима (p_durable [6.3]),
3. Доля занятых в сельском хозяйстве (cnts_percent_work_force_in_agriculture [5.5]),
4. Население (cnts_pop1 [4.6]),
5. Доля занятых в сфере услуг (cnts_percent_work_force_in_other_activity [4.3]).

Вторая группа факторов (с коэффициентом importance > 1,5): Площадь (cnts_area1), Доля занятых в промышленности (cnts_percent_work_force_in_industry), Число поступивших в вузы, на 1000 чел. (cnts_university_enrollment), Цена на нефть марки Brent в номинальных ценах (Brent_price_nominal), Оценка Polity IV (p_polity), Возраст государственности, независимость (nationality_sovereignty), Индекс цен на товарные продукты питания (commodity_food), Суммарный индекс свобод (fh_status_sum).

Данные переменные совместно отвечают за 63 % общей значимости модели. Вторая группа факторов отвечает за 16 %, третья за 15 % общей объясняющей способности модели.

В результате оценки моделей для переменной «**Массовые беспорядки**» (cnts_domestic6) были выделены следующие факторы:

1. Население (cnts_pop1 [16]).
2. Индекс потребительских цен (consumer_price_index [2.5]).
3. Доля занятых в промышленности (cnts_percent_work_force_in_industry [2.5]).
4. Индекс регулирования участия в политике (p_parreg [2.2]).
5. Долговечность режима (p_durable [2.1]).

Вторая по важности группа переменных (с коэффициентом importance > 1,5): Цена на золото (Gold_price), Плотность населения (cnts_pop_density), Доля грамотных (cnts_percent_literate).

Все эти группы совместно отвечают за 46 % общей значимости модели. Вторая группа факторов отвечает за 5 % общей важности модели, третья группа за 16 % зависимой переменной «Массовые беспорядки».

Тренировка 100 бустинговых моделей кросс-валидации позволяет выделить следующие основные независимые факторы влияющих на переменную «**Государственные перевороты и попытки переворотов**» (cnts_domestic7):

1. Долговечность режима (p_durable [9.5]).
2. ВВП на душу населения по ППС, в постоянных ценах 2011 г. (gdp_per_capita_PPP_Mad [3.5]).
3. Индекс регулирования участия в политике (p_parreg [3]).
4. Эффективность законодательной власти (cnts_effectiveness_of_legislature [2.9]),
5. Оценка Polity IV (p_polity [2]).

Вторая группа факторов: Темпы роста населения (cnts_pop_growth), Индекс конкурентоспособности участия (p_parcomp), Плотность населения (cnts_pop_density), Государственный долг, в текущей национальной валюте (central_gov_debt_current_LCU), Индекс фракционности (cnts_polit01), Доля грамотных (cnts_percent_literate), Индекс регулирования главного исполнительного органа (p_xrreg), Площадь (cnts_area1), Доля валовых внутренних инвестиций в ВВП (gross_fixed_capital_formation_percent_gdp).

Приведенные факторы совместно объясняют 50 % общей значимости модели. Вторая группа факторов объясняет 15 %, третья – 14 %.

По результатам тренировки моделей для зависимой переменной «**Антиправительственные демонстрации**» (cnts_domestic8) мы выделили следующие наиболее важные факторы:

1. Население (cnts_pop1 [14.2]).
2. Число поступивших в вузы на 1000 человек (cnts_university_enrollment [6]),
3. Индекс потребительских цен (consumer_price_index [3.3]),
4. Плотность населения (cnts_pop_density [2.6]),
5. Долговечность режима (p_durable [2.5]).

Вторая группа факторов (с коэффициентом importance > 1,5): ВВП по ППС, в постоянных долларах США 2011 г. (gdp_PPP), Доля валовых внутренних инвестиций в ВВП (gross_fixed_capital_formation_percent_gdp), Темпы роста ВВП по ППС (gdp_PPP_annual_growth).

Данные переменные совместно отвечают за 43 % общей значимости модели. Вторая группа факторов отвечает за 6 %, третья за 9 % общей объясняющей способности модели.

По результатам тренировки моделей для зависимой переменной «**Индекс социально-политической дестабилизации**» (cnts_domestic9) мы выделили следующие наиболее важные факторы:

1. Население (cnts_pop1 [18.1]),
2. Долговечность режима (p_durable [5.6]),
3. Индекс регулирования участия в политике (p_parreg [2.9]),
4. Плотность населения (cnts_pop_density [2]),
5. Экспорт на душу населения, в долларах (cnts_exports_per_capita [1.7]).

Вторая группа факторов (с коэффициентом importance > 1,5): Население (population), Индекс потребительских цен (consumer_price_index), Доля занятых в промышленности (cnts_percent_work_force_in_industry).

В третью по важности группу входят следующие 11 факторов: Возраст государственности, независимость (nationality_sovereignty), Доля грамотных (cnts_percent_literate), Темпы роста ВВП по ППС (gdp_PPP_annual_growth), Индекс фракционности (cnts_polit01), Доля валовых внутренних инвестиций в ВВП (gross_fixed_capital_formation_percent_gdp), Количество городов на душу населения (cnts_population_cities_over_per_capita), Цена на нефть марки Brent в номинальных ценах (Brent_price_nominal), Цена на золото (Gold_price), Темпы роста ВВП на душу населения по ППС, в постоянных ценах 2011 г. (growth_rate_gdp_per_capita_PPP_WB), Це-

на на серебро (Silver_price), Темпы роста ПИИ, приток (fdi_ inward_percent_gdp).

Данные переменные совместно отвечают за 48 % общей значимости модели. Вторая группа факторов отвечает за 5 %, третья за 13 % общей объясняющей способности модели.

В результате оценки моделей для переменной «**Индекс социально-политической дестабилизации (опережающая на 1 год)**» (cnts_domestic9_next) были выделены следующие факторы:

1. Население (cnts_pop1 [16.2]).
2. Население (population [3.9]).
3. Долговечность режима (p_durable [3.5]).
4. Индекс регулирования участия в политике (p_parreg [3.2]).
5. Количество городов на душу населения (cnts_population_cities_over_per_capita [2.4]).

Вторая по важности группа переменные (с коэффициентом importance > 1,5): Доля занятых в промышленности (cnts_percent_ work_force_in_industry), Плотность населения (cnts_pop_density).

Все эти группы совместно отвечают за 43 % общей значимости модели. Вторая группа факторов отвечает за 4 % общей важности модели, третья группа за 10 %.

По результатам тренировки моделей для зависимой переменной «**Индекс социально-политической дестабилизации (с лагом в +1 год)**» (cnts_domestic9_prev) мы выделили следующие наиболее важные факторы:

1. Население (cnts_pop1 [15.9]).
2. Долговечность режима (p_durable [7]).
3. Индекс регулирования участия в политике (p_parreg [3]).
4. Доля занятых в промышленности (cnts_percent_work_force_ in_industry [2.4]).
5. Доля валовых внутренних инвестиций в ВВП (gross_fixed_ capital_formation_percent_gdp [2]).

Вторая группа факторов (с коэффициентом importance > 1,5): Плотность населения (cnts_pop_density), Экспорт на душу населения, в долларах (cnts_exports_per_capita), Население (population), Индекс фракционности (cnts_polit01).

Данные переменные совместно отвечают за 48 % общей значимости модели. Вторая группа факторов отвечает за 7 %, третья за 11 % общей объясняющей способности модели.

По результатам тренировки моделей для переменной «**Государственные перевороты и попытки переворотов**» (coup_detat_

failed_coup_detat) мы выделили следующие наиболее важные независимые факторы:

1. Комбинированная оценка Polity IV (p_polity_2_2 [14.7]).
2. Эффективность законодательной власти (cnts_effectiveness_of_legislature [9.6]).
3. Тип режима (cnts_type_regime [7.7]).
4. Долговечность режима (p_durable [4.8]).
5. Индекс регулирования участия в политике (p_parreg [3.1]).

Во вторую по важности группу (с коэффициентом importance > > 1,5) входят следующие 4 переменных: Коэффициент насилия, связанного со сменой режима (MAGVIOL), Тим смены режима (POLITYX), Экспорт на душу населения, в долларах (cnts_exports_per_capita).

Данные переменные совместно отвечают за 58 % общей значимости модели. Вторая группа факторов отвечает за 8 %, третья за 10 % общей объясняющей способности модели.

По результатам тренировки моделей для переменной «**Количество террористических атак**» (n_terror_attack) мы выделили следующие наиболее важные независимые факторы:

1. Городское население в возрасте от 15 до 24 лет (urban_15__24 [6.2]).
2. Плотность населения (cnts_pop_density [3]).
3. Доля сельского хозяйства в ВВП (share_agriculture [2.9]).
4. ВВП, в текущей национальной валюте (gdp_current_LCU [2.6]).
5. Городское население в возрасте от 20 до 29 лет (urban_20__29 [2.5]).

Во вторую по важности группу (с коэффициентом importance > 1,5) входят следующие 8 переменных: Доля валовых внутренних инвестиций в ВВП (gross_fixed_capital_formation_percent_gdp), Городское население в возрасте от 15 до 29 лет (urban_15__29), Дельта городского населения (delta_urban_population_UN), Фракционность элит (factionalized_elites), Возраст государственности, независимость (nationality_sovereignty), Комбинированная оценка Polity IV (p_polity_2_2), Индекс цен на товарные продукты питания и напитки (commodity_food_beverage), Население в возрасте от 0 до 4 лет (за 5 лет) (X0__4).

Данные переменные совместно отвечают за 49 % общей значимости модели. Вторая группа факторов отвечает за 16 %, третья за 16 % общей объясняющей способности модели.

По результатам тренировки моделей для переменной «**Количество убитых в терактах**» (nkill) мы выделили следующие наиболее важные независимые факторы:

1. Население в возрасте от 0 до 4 лет (за 5 лет) (X0__4 [9.6]).
2. Коэффициент смертности (mort_rate [4.1]).
3. Цена на нефть марки Brent в номинальных ценах (Brent_price_nominal [2.5]),
4. ВВП, в текущей национальной валюте (gdp_current_LCU [2.1]).
5. Доля сельского хозяйства в ВВП (share_agriculture [2.1]).

Во вторую по важности группу (с коэффициентом importance > > 1,5) входят следующие 6 переменных: Темпы роста ПИИ, приток (fdi_inward_percent_gdp), Доля городского населения (share_urban_population_UN), NA (share_15_24), Городское население в возрасте от 15 до 24 лет (urban_15__24), Плотность населения (cnts_pop_density), Темпы роста ПИИ, отток (fdi_outward_percent_gdp).

Данные переменные совместно отвечают за 47 % общей значимости модели. Вторая группа факторов отвечает за 10 %, третья за 16 % общей объясняющей способности модели.

Анализ робастности результатов

Использование подхода k-fold кросс-валидации позволяет нам контролировать объясняющие способности модели для всего представленного набора данных и также позволяет нам оценить устойчивость полученных результатов к случайным флуктуациям или надуманным зависимостям. Мы выбрали общее число кросс-валидации на одну модель равное 100 для обеспечения сходимости результатов к общей закономерности. Пример оценки отдельных моделей показывает, что значимость отдельных факторов и значимость модели в целом может варьироваться от подвыборки, на которой строится оценка. С увеличением числа модельных оценок формируется устойчивая картина. Использование 100 кросс-валидаций дает стабильные ранги (по важности) для первых 5–10 переменных и выделяет группу чуть менее важных основных переменных, которые, тем не менее, имеют высокую важность для объяснения зависимой переменной.

Общая оценка модели (тестовые ошибки и AUC) при таком количестве валидаций, также оказывается устойчивой. В Приложении 3 приведены результаты качества оценки моделей градиентного бустинга для альтернативных 100 кросс-валидаций, то есть с

другим разбиением на тестовые и обучающие выборки, отличные от тех, что приводятся в данной работе (для базовых 100 кросс-валидаций; аналогичные результаты приведены выше в Табл. 3). Из сопоставления Приложения 3 с Табл. 3 видно, что статистика по базовым 100 кросс-валидациям (каждой из 14 зависимых переменных) и альтернативные 100 кросс-валидаций отличаются незначимо (в пределах оцененной дисперсии).

В Приложении 4 приведены общие оценки важности основных переменных, полученные для альтернативных 100 кросс-валидациях.

Мы не стали обобщать альтернативные и базовые кросс-валидации для оценки значимости переменных, так как в дальнейшее увеличение числа кросс-валидации не изменяет полученные результаты.

Заключение и обсуждение результатов

Используя методы градиентного бустинга, мы отобрали набор наиболее важных переменных для мониторинга важных показателей социально-политической дестабилизации.

Среди наиболее важных переменных выделяются переменные, которые описывают историю существования и устойчивости режима (Долговечность режима, Возраст государственности[2]), переменные характеризующие тип режима (Индекс регулирования участия в политике, Комбинированная оценка Polity IV), переменные, характеризующие структуру населения и занятости (Население в возрасте от 0 до 4 лет (за 5 лет), Доля занятых в промышленности), переменные отражающие состояние мировой конъюнктуры (Цена на золото, Индекс потребительских цен).

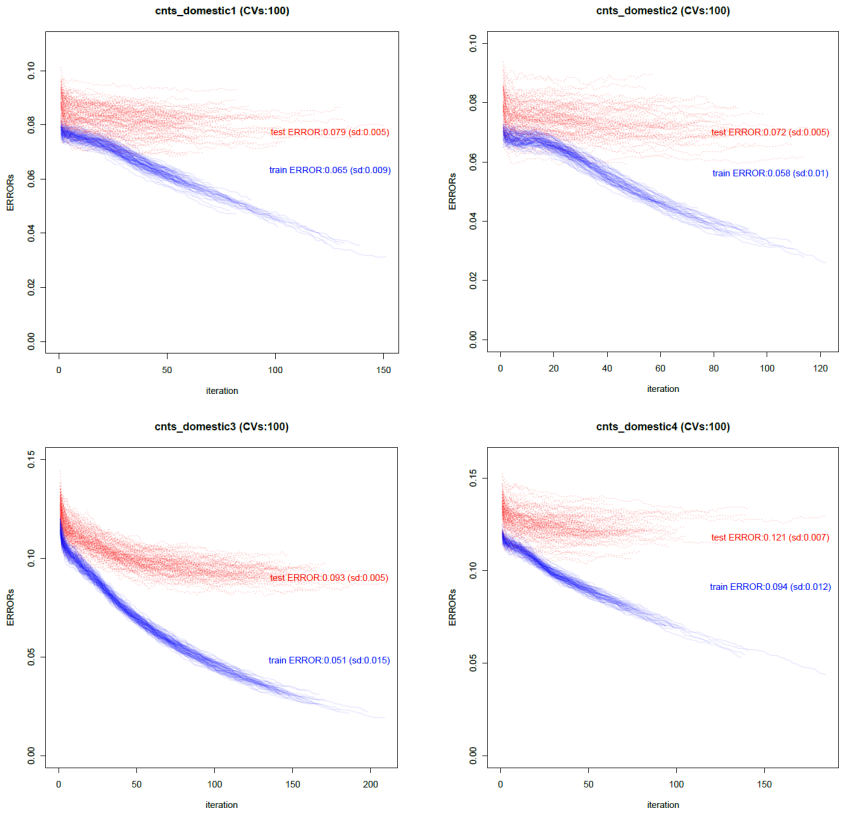
В текущей версии мы использовали модели, в которых анализирует ограниченное число аспектов нестабильности (измеряемый с помощью индексов нестабильности CNTS и еще 3 дополнительных переменных). Дальнейшее уточнение, формализация понятия нестабильности поможет уточнить набор переменных для других измерений нестабильности.

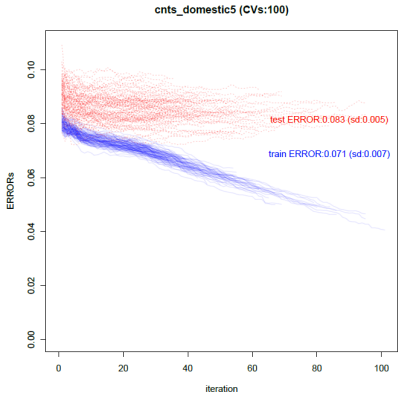
Мы использовали модель «common pool» для анализа наиболее важных факторов, влияющих на нестабильность, построение моделей, направленных на предсказание нестабильности, также позволит в дальнейшем уточнить набор наиболее важных факторов.

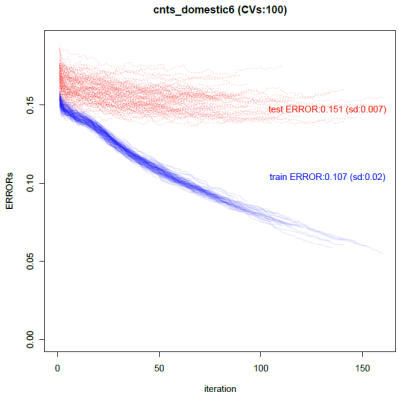
---

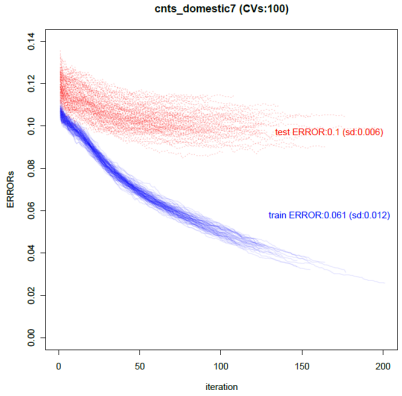
[2] Как уже отмечалось, это хорошо согласуется с результатами, опубликованными в предыдущем выпуске Мониторинга (Гринин и др. 2017).

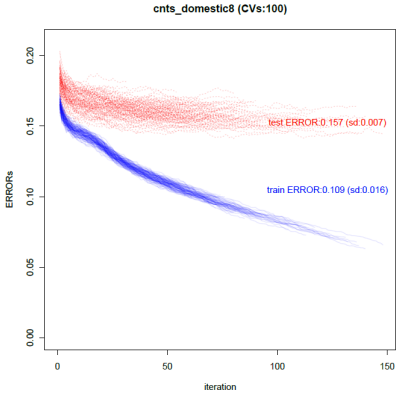
Приложение 1. Графики динамики ошибок для обучающих и тестовых выборок

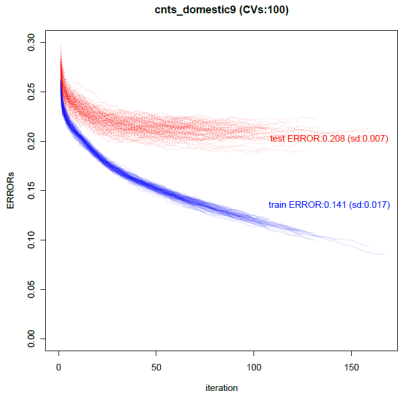


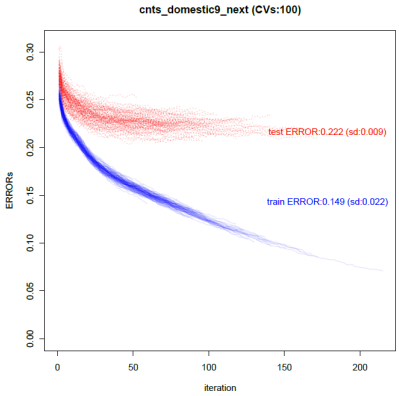

cnts_domestic5 (CVs:100)
test ERROR:0.083 (sd:0.005)
train ERROR:0.071 (sd:0.007)
ERRORs
iteration

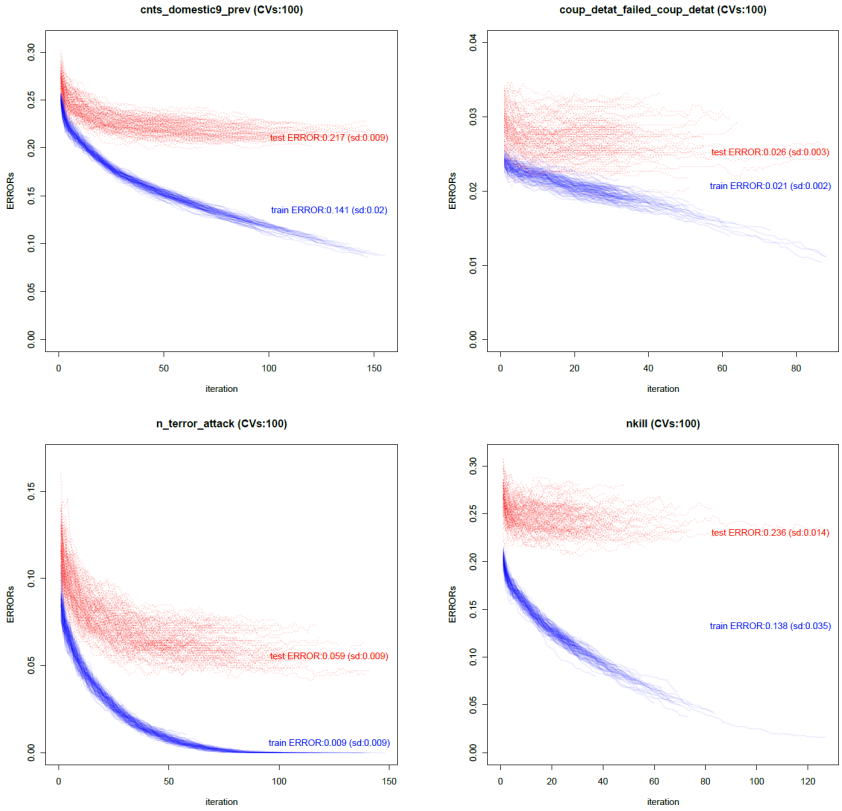

cnts_domestic6 (CVs:100)
test ERROR:0.151 (sd:0.007)
train ERROR:0.107 (sd:0.02)
ERRORs
iteration


cnts_domestic7 (CVs:100)
test ERROR:0.1 (sd:0.006)
train ERROR:0.061 (sd:0.012)
ERRORs
iteration


cnts_domestic8 (CVs:100)
test ERROR:0.157 (sd:0.007)
train ERROR:0.109 (sd:0.016)
ERRORs
iteration

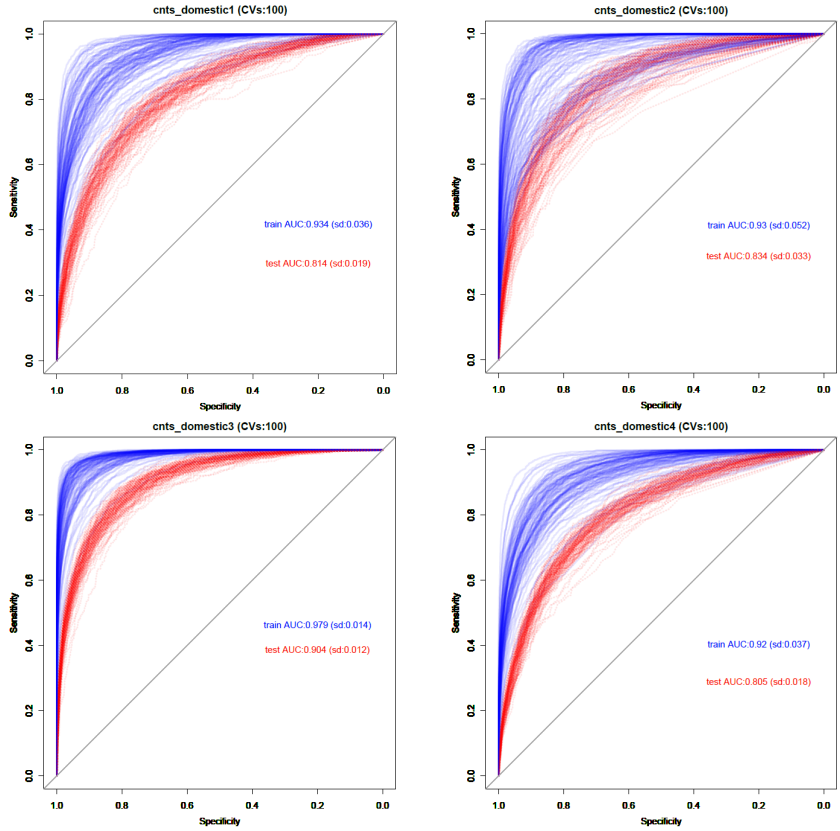

cnts_domestic9 (CVs:100)
test ERROR:0.208 (sd:0.007)
train ERROR:0.141 (sd:0.017)
ERRORs
iteration

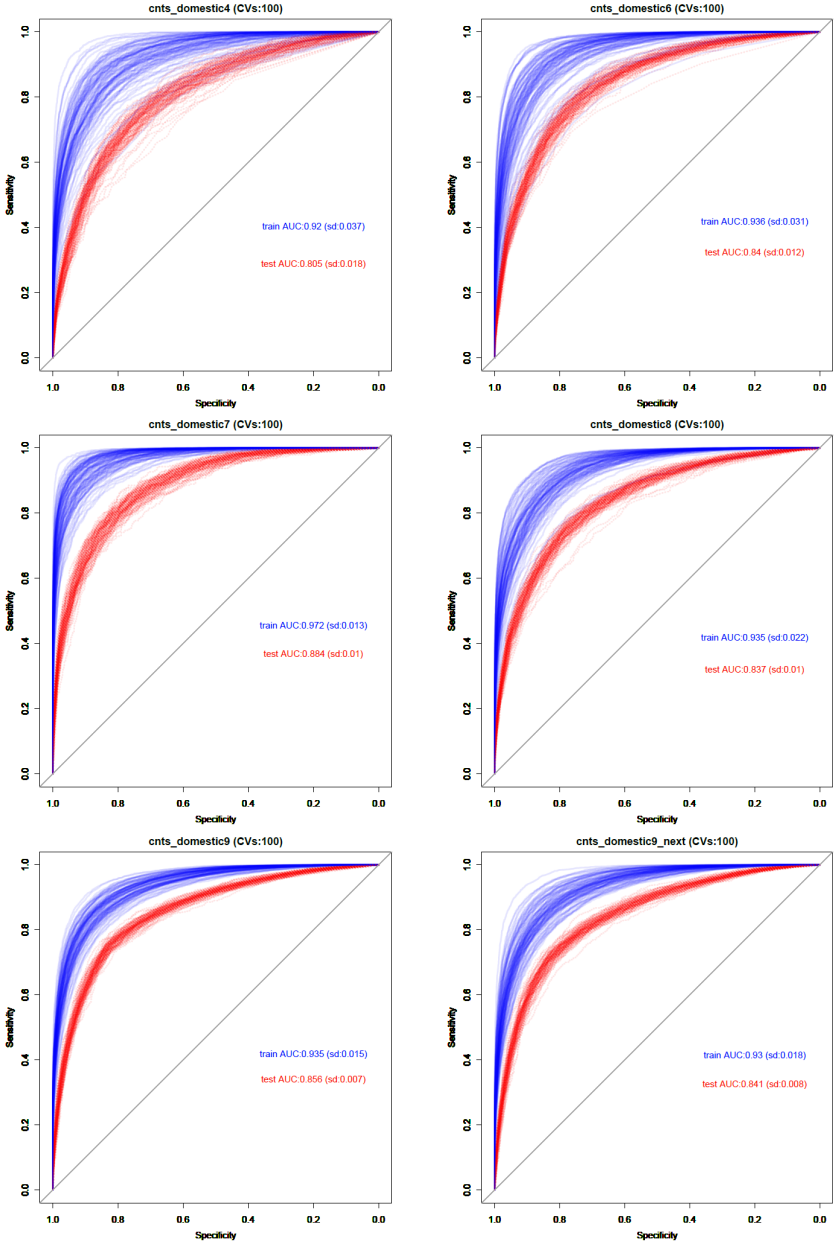

cnts_domestic9_next (CVs:100)
test ERROR:0.222 (sd:0.009)
train ERROR:0.149 (sd:0.022)
ERRORs
iteration

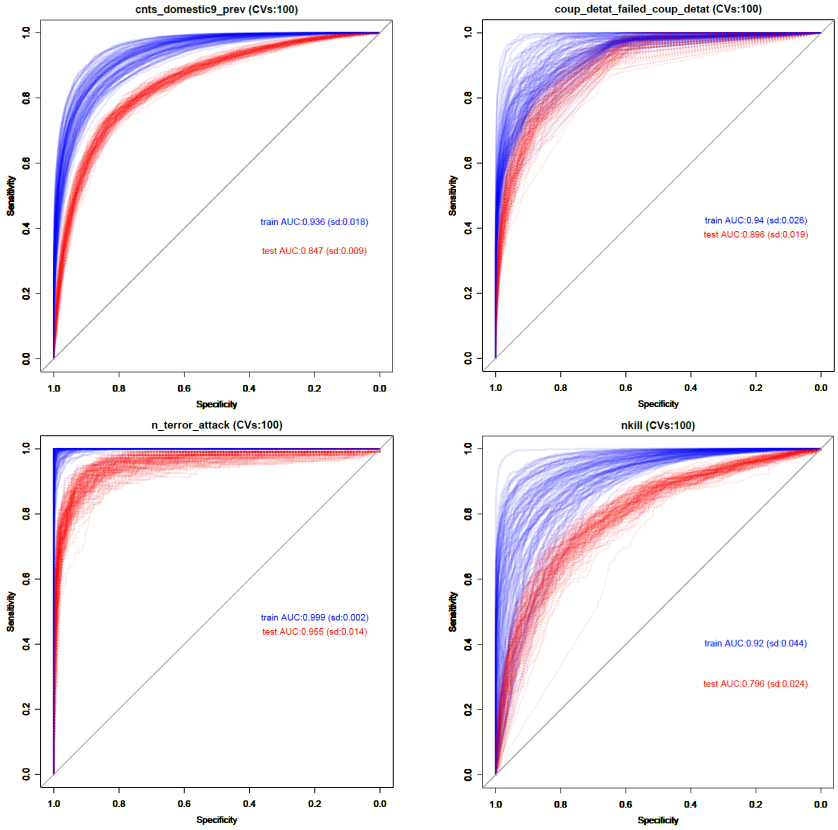


Примечание: на всех графиках приводится среднее значение и стандартное отклонение (в скобках) значения ошибок рассчитанное для 100 результирующих оценок.

Приложение 2. ROC-кривые для обучающих и тестовых выборок




cnts_domestic4 (CVs:100)
train AUC:0.92 (sd:0.037)
test AUC:0.805 (sd:0.018)
cnts_domestic6 (CVs:100)
train AUC:0.936 (sd:0.031)
test AUC:0.84 (sd:0.012)
cnts_domestic7 (CVs:100)
train AUC:0.972 (sd:0.013)
test AUC:0.884 (sd:0.01)
cnts_domestic8 (CVs:100)
train AUC:0.935 (sd:0.022)
test AUC:0.837 (sd:0.01)
cnts_domestic9 (CVs:100)
train AUC:0.935 (sd:0.015)
test AUC:0.856 (sd:0.007)
cnts_domestic9_next (CVs:100)
train AUC:0.93 (sd:0.018)
test AUC:0.841 (sd:0.008)
Sensitivity
Specificity



Примечание: на всех графиках приводится среднее значение и стандартное отклонение (в скобках) показателя AUC рассчитанное для 100 прогнозных оценок.

Приложение 3. Результаты оценки моделей нестабильности для 1400 базовых и 1400 альтернативных наборов кросс-валидации

Табл. П3.1 Альтернативные 1400 кросс-валидаций: Оценки качества моделей на обучающих и тестовых выборках

Переменная	Обучающая выборка (train)		Тестовая выборка (test)	
	Error	AUC	Error	AUC
cnts_domestic1	0.066 (0.009)	0.927 (0.045)	0.080 (0.005)	0.812 (0.020)
cnts_domestic2	0.059 (0.010)	0.924 (0.055)	0.072 (0.004)	0.833 (0.031)
cnts_domestic3	0.051 (0.013)	0.979 (0.011)	0.093 (0.005)	0.904 (0.009)
cnts_domestic4	0.095 (0.012)	0.916 (0.038)	0.121 (0.006)	0.804 (0.018)
cnts_domestic5	0.071 (0.007)	0.927 (0.032)	0.083 (0.005)	0.872 (0.020)
cnts_domestic6	0.110 (0.019)	0.932 (0.030)	0.152 (0.007)	0.839 (0.012)
cnts_domestic7	0.059 (0.012)	0.973 (0.013)	0.099 (0.005)	0.884 (0.009)
cnts_domestic8	0.109 (0.017)	0.934 (0.023)	0.157 (0.007)	0.837 (0.008)
cnts_domestic9	0.140 (0.017)	0.936 (0.015)	0.208 (0.007)	0.856 (0.006)
cnts_domestic9_next	0.148 (0.020)	0.931 (0.018)	0.223 (0.008)	0.841 (0.008)
cnts_domestic9_prev	0.143 (0.022)	0.934 (0.018)	0.217 (0.008)	0.847 (0.008)
coup_detat_failed_coup_detat	0.021 (0.002)	0.945 (0.026)	0.026 (0.002)	0.897 (0.019)
n_terror_attack	0.007 (0.007)	0.999 (0.001)	0.059 (0.009)	0.955 (0.013)
nkill	0.136 (0.034)	0.925 (0.041)	0.238 (0.014)	0.796 (0.020)

Примечание: приводится среднее значение и стандартное отклонение (в скобках) ошибок и показателя AUC рассчитанное для альтернативных (отличных от базовых) 100 прогнозных оценок для каждой переменный.

Таблица П3.2. повторяет таблицу 3 и проводится для сопоставления с таблицей П3.1

Табл. П3.2 Базовые 1400 кросс-валидаций: Оценки качества моделей на обучающих и тестовых выборках

Переменная	Обучающая выборка (train)		Тестовая выборка (test)	
	Error	AUC	Error	AUC
cnts_domestic1	0,065 (0,009)	0,934 (0,036)	0,079 (0,005)	0,814 (0,019)
cnts_domestic2	0,058 (0,010)	0,930 (0,052)	0,072 (0,005)	0,834 (0,033)
cnts_domestic3	0,051 (0,015)	0,979 (0,014)	0,093 (0,005)	0,904 (0,012)
cnts_domestic4	0,094 (0,012)	0,920 (0,037)	0,121 (0,007)	0,805 (0,018)
cnts_domestic5	0,071 (0,007)	0,927 (0,033)	0,083 (0,005)	0,871 (0,022)
cnts_domestic6	0,107 (0,020)	0,936 (0,031)	0,151 (0,007)	0,840 (0,012)
cnts_domestic7	0,061 (0,012)	0,972 (0,013)	0,100 (0,006)	0,884 (0,010)
cnts_domestic8	0,109 (0,016)	0,935 (0,022)	0,157 (0,007)	0,837 (0,010)
cnts_domestic9	0,141 (0,017)	0,935 (0,015)	0,208 (0,007)	0,856 (0,007)
cnts_domestic9_next	0,149 (0,022)	0,930 (0,018)	0,222 (0,009)	0,841 (0,008)
cnts_domestic9_prev	0,141 (0,020)	0,936 (0,018)	0,217 (0,009)	0,847 (0,009)
coup_detat_failed_coup_detat	0,021 (0,002)	0,940 (0,026)	0,026 (0,003)	0,896 (0,019)
n_terror_attack	0,009 (0,009)	0,999 (0,002)	0,059 (0,009)	0,955 (0,014)
nkill	0,138 (0,035)	0,920 (0,044)	0,236 (0,014)	0,796 (0,024)

Примечание: приводится среднее значение и стандартное отклонение (в скобках) ошибок и показателя AUC рассчитанное для основных 100 прогнозных оценок по каждой переменной.

Приложение 4. Результаты оценки общей важности переменных полученные для альтернативных 1400 наборов кросс-валидации

Тренировка бустинговых моделей на 100 альтернативных кросс-валидациях (отличных от базовых, результаты по которым анализируются в статье) позволяет выделить следующие важнейшие факторы:

1. Население (cnts_pop1 [8.7]).
2. Долговечность режима (p_durable [4.2]).
3. Индекс регулирования участия в политике (p_parreg [2.3]).
4. Плотность населения (cnts_pop_density [1.9]).
5. Доля занятых в промышленности (cnts_percent_work_force_in_industry [1.5]).
6. Комбинированная оценка Polity IV (p_polity_2_2 [1.5]).
7. Возраст государственности, независимость (nationality_sovereignty [1.5]).
8. Население в возрасте от 0 до 4 лет (за 5 лет) (X0__4 [1.4]).
9. Индекс потребительских цен (consumer_price_index [1.3]).
10. Экспорт на душу населения, в долларах (cnts_exports_per_capita [1.3]).

В следующую по важности группу (при альтернативных кросс-валидациях) вошли 15 факторов: Доля валовых внутренних инвестиций в ВВП (gross_fixed_capital_formation_percent_gdp), Количество городов на душу населения (cnts_population_cities_over_per_capita), Цена на золото (Gold_price), Население (population), Темпы роста ВВП по ППС (gdp_PPP_annual_growth), Площадь (cnts_area1), Индекс фракционности (cnts_polit01), ВВП, в текущей национальной валюте (gdp_current_LCU), Доля грамотных (cnts_percent_literate), Темпы роста ВВП на душу населения по ППС, в постоянных ценах 2011 г. (growth_rate_gdp_per_capita_PPP_WB), Темпы роста ПИИ, приток (fdi_inward_percent_gdp), Эффективность закондательной власти (cnts_effectiveness_of_legislature), Цена на нефть марки Brent в номинальных ценах (Brent_price_nominal), Число поступивших в вузы, на 1000 чел. (cnts_university_enrollment), Темпы роста населения (cnts_pop_growth).

Все перечисленные факторы совместно объясняют 42 % общей значимости модели, а последние 15 факторов 17 %.

Библиография

Васькин И. А., Цирель С. В., Коротаев А. В. 2018. Экономический рост, образование и терроризм: опыт количественного анализа. *Социологический журнал* 24(2): 28–65.

Гринин Л. Е., Билюга С. Э., Коротаев А. В., Гринин А. Л. 2017. Возраст государства и социально-политическая дестабилизация: предва-

рительные результаты количественного анализа. *Системный мониторинг глобальных и региональных рисков* 8: 141–169.

Коротаев А. В., Билюга С. Э., Шишкина А. Р. 2016. ВВП на душу населения, уровень протестной активности и тип режима: опыт количественного анализа. *Сравнительная политика* 4(26): 72–94.

Коротаев А., Васькин И., Билюга С. 2017. Гипотеза Олсона – Хантингтона о криволинейной зависимости между уровнем экономического развития и социально-политической дестабилизацией: опыт количественного анализа. *Социологическое обозрение* 16(1): 9–49.

Коротаев А. В., Зинькина Ю. В. 2011. Египетская революция 2011 года: социодемографический анализ. *Историческая психология и социология истории* 4(2): 5–29.

Коротаев А. В., Зинькина Ю. В. 2012. Структурно-демографические факторы «арабской весны». *Протестные движения в арабских странах. Предпосылки, особенности, перспективы* / Ред. И. В. Следзевский, А. Д. Саватеев. М.: Либроком/URSS. С. 28–40.

Малков С. Ю., Коротаев А. В., Исаев Л. М., Кузьминова Е. В. 2013. О методике оценки текущего состояния и прогноза социальной нестабильности: опыт количественного анализа событий Арабской весны. *Полис. Политические исследования* 4: 137–162.

Цирель С. В. 2012. Условия возникновения революционных ситуаций в арабских странах. Арабская весна 2011 года. *Системный мониторинг глобальных и региональных рисков* / Ред. А. В. Коротаев, Ю. В. Зинькина, А. С. Ходунов. М.: ЛИБРОКОМ/URSS. С. 162–173.

Цирель С. В. 2015. К истокам украинских революционных событий 2013–2014 гг. *Системный мониторинг глобальных и региональных рисков* / Ред. Л. Е. Гринин, А. В. Коротаев, Л. М. Исаев, А. Р. Шишкина. Волгоград: Учитель. С. 57-83.

Astashkin A., Chuvilin K. 2017. Syntax description synthesis using gradient boosted trees. *Open Innovations Association (FRUCT), 2017 20th Conference of*. IEEE. Pp. 32–39/

Banks A. S., Wilson K. A. 2018. *Cross-National Time-Series Data Archive.* Jerusalem, Israel. Databanks International. URL: https://www.cntsdata.com/.

Breiman L. 2001. Random forests. *Machine learning* 45(1): 5–32.

Chen T., Guestrin C. 2016. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785–794). ACM.

Connelly R., C. J. Playford V. Gayle and C. Dibben. 2016. 'The Role of Administrative Data in the Big Data Revolution in Social Science Research'. *Social Science Research* 59: 1–12.

Coppock A., Guess A., Ternovski J. 2016. 'When Treatments are Tweets: A Network Mobilization Experiment over Twitter', Political Behavior, 38(1): 105–128.

Donnay K. 2017. Big Data for Monitoring Political Instability. *International Development Policy| Revue internationale de politique de développement*. T. 8. №. 8.1.

Donnay K., E., Dunford E. C., McGrath, D. Backer and D. E. Cunningham. 2016. 'MELTT: Matching Event Data by Location, Time and Type', paper presented at the Annual Conference of the Midwest Political Science Association, Chicago, April 2016.

Esty, D., Goldstone, J. A., Gurr, T. R., Harff, B., Levy, M., Dabelko, G. D., Surko, P., Unger, A. N. 1998. State Failure Task Force Report: Phase II Findings. McLean, VA: Sci. Appl. Int. Corp. Failed and Fragile States. 2014. URL: http://www4.carleton.ca/cifp/ (accessed 20.02.2015).

Freedom in the World. 2017. *Freedom House*. URL: https://freedomhouse.org/report/methodology-freedom-world–2017 (дата обращения: 02.02. 2015).

Friedman J. H. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics* 29(5): 1189–1232.

Korotayev A., Bilyuga S., Shishkina A. 2018. GDP Per Capita and Protest Activity: A Quantitative Reanalysis. *Cross-Cultural Research* 52(4): 406–440.

Korotayev A., Shulgin S. and Zinkina J. 2017. *Country Risk Analysis Based on Demographic and Socio-Economic Data*. Moscow: RANEPA. URL: http://dx.doi.org/10.2139/ssrn.2944064

Korotayev A., Zinkina J., Kobzeva S., Bogevolnov J., Khaltourina D., Malkov A., Malkov S. 2011. A Trap at the Escape from the Trap? Demographic-Structural Factors of Political Instability in Modern Africa and West Asia. *Cliodynamics* 2(2): 276–303.

Kuhn M. 2008. Building Predictive Models in R Using the caret Package. *Journal of Statistical Software* 28(5): 1–26.

Maksims Volkovs, Guang Wei Yu, Poutanen T. 2017. Content-based Neighbor Models for Cold Start in Recommender Systems. In Proceedings of RecSys Challenge '17, Como, Italy, August 27, 2017, 6 pages. https://doi.org/10.1145/3124791.3124792

Marshall M. G. 2016. *Coup D'État Events, 1946–2015 Codebook*. Vienna, VA: Center for Systemic Peace.

Mironov V., Guschin A. 2015. 1st place of the CERN LHCb experiment Flavour of Physics competition, URL: http://blog.kaggle.com/2015/11/30/flavour-of-physics-technical-write-up-1st-place-go-polar-bears/.

Polity IV. 2015. *Annual Time-Series, 1800–2015*. URL: http://www.systemicpeace.org/polity/polity4.htm (дата обращения: 02.02.2015).

Przeworski A., Alvarez M. E., Cheibub J. A. Limongi F. 2000. Adam Przeworski, ed. Democracy and Development; Political Institutions and Well-Being in the World, 1950–1990. New York: Cambridge University Press.

Sandulescu V., Chiru M. 2016. Predicting the future relevance of research institutions-The winning solution of the KDD Cup 2016. *arXiv preprint arXiv:1609.02728*.

Slinko E., Bilyuga S., Zinkina J., Korotayev A. 2017. Regime type and political destabilization in cross-national perspective: A re-analysis. *Cross-Cultural Research* 51(1): 26-50.

START [National Consortium for the Study of Terrorism and Responses to Terrorism]. **2016.** *Global Terrorism Database*. College Park, MD: National Consortium for the Study of Terrorism and Responses to Terrorism. https://www.start.umd.edu/gtd/ URL:

Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, Kailong Chen, Rory Mitchell, Ignacio Cano, Tianyi Zhou, Mu Li, Junyuan Xie, Min Lin, Yifeng Geng and Yutian Li. 2018. xgboost: Extreme Gradient Boosting. R package version 0.71.2. https://CRAN.R-project.org/package=xgboost

Volkovs M., Yu G. W., Poutanen T. 2017. Content-based Neighbor Models for Cold Start in Recommender Systems. In *Proceedings of the Recommender Systems Challenge* (p. 7). ACM.