АНАЛИЗ ФАКТОРОВ СОЦИАЛЬНО-ПОЛИТИЧЕСКОЙ НЕСТАБИЛЬНОСТИ В СТРАНАХ АФРАЗИЙСКОЙ МАКРОЗОНЫ С ПОМОЩЬЮ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ*

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

С использованием методов машинного обучения мы анализировали различные группы факторов, влияющих на социально-политическую нестабильность. Анализ проводился на основе обновленной расширенной базы данных, включающей более 300 независимых переменных, для 19 тыс. наблюдений в формате страно-год. В качестве источника данных по социально-политической нестабильности используются данные The Cross-National Time Series (CNTS) и Global Terrorism Database (GTB). Данные по релевантным независимым переменным мы брали из разных источников, таких как World Bank, United Nation Population Division, Polity IV, Maddison Database, Worldwide Governance Indicators, CNTS, World Value Survey и др. Анализ факторов проводили для отдельных макрорегионов: Афразийской макрозоны; Латинской Америки; Восточной Европы, стран Африки южнее Сахеля и подгруппы наиболее развитых западных стран. Были проанализированы и систематизированы факторы, наиболее важные для оценки социально-политической нестабильности для каждого из макрорегионов.

Для Афразийской макрозоны нестабильности в качестве наиболее важных выделяются несколько групп факторов: описывающие историю режима, описывающие экономику, описывающие политический строй. Наиболее важными оказываются факторы, описывающие историю политической (не)стабильности (число по-

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

пыток переворотов за предшествующие 5 лет; долговечность режима; сколько лет текущий глава государства занимает свой пост). Среди экономических факторов наиболее важные: темпы роста ВВП; ВВП, в текущей национальной валюте; индекс потребительских цен; темпы роста населения; индекс глобализации; доля ПИИ в ВВП. Среди факторов, описывающих тип режима, наиболее важными оказываются: финансирование программ со стороны Америки, в долларах; индекс либеральной демократии; индекс регулирования участия. Отдельно выделяются факторы, описывающие размер экономики: численность населения, численность городского населения; население в возрасте от 0 до 4 лет (за 5 лет).

В отличие от стран Афразийской макрозоны для группы стран Латинской Америки дополнительно среди наиболее важных факторов, определяющих различные формы социально-политической нестабильности, появляется фактор рыночной конъюнктуры: цена на нефть марки Brent в номинальных ценах и такие характеристики, как индекс политической фракционности и доля занятых в сфере услуг. Среди наиболее значимых отсутствуют такие факторы, как индекс глобализации.

Для группы стран Восточной Европы в отличие от стран Афразийской макрозоны в качестве наиболее важных факторов добавляются доля занятых в промышленности, экспорт на душу населения, а также степень коррумпированности режима. Как и для стран Латинской Америки, важным фактором выступает цена на нефть марки Brent в номинальных ценах.

Традиционный подход к анализу факторов, влияющих на социально-политическую нестабильность, – это определение экспертами факторов, которые потенциально могут оказывать влияние на различные социально-политические процессы. Такой экспертный выбор делается на основании теоретических или эмпирических соображений.

Среди факторов, которые могут оказывать влияние на социально-политические процессы, можно выделить несколько групп: социально-демографические характеристики общества; экономические факторы; факторы политического устройства и государствен-

ного управления; факторы, описывающие историю режима, его устойчивость, и другие.

Среди экономических факторов 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.

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

Характеристики, описывающие модель социального и политического устройства: тип политического режима, тип правовой системы, характеристики формальных и неформальных институтов, характеристики системы государственного управления. Выделяются географические и исторические факторы: описывающие положение, взаимное расположение, совместное историческое.

Подобная классификация значимых факторов не будет законченной, так как всегда есть параметры, которые можно отнести к разным группам факторов или выделить в отдельную, также появляются новые характеристики общества, которые оказывают значимое воздействие на социально-политические процессы, например, связанные с распространением новых технологий, новых устройств, новых сервисов и программных продуктов.

Обзор отдельных подходов к анализу нестабильности приведен в нашей предыдущей работе (см.: Korotaev *et al*. 2017). Анализ отдельных факторов представлен, например, в работах, как отечественных, так и зарубежных авторов (Esty *et al*. 1998; Цирель 2012; 2015; Коротаев, Зинькина 2012; Малков и др. 2013; Przeworski *et al*. 2000) и множестве других.

Анализ социально-политической нестабильности можно проводить и с использованием микроданных (отдельных событий, высокой географической детализации, поведения отдельных людей и т. п.) и для работы, с которыми используются класс модели условно называемых методами машинного обучения. Обзор подходов представлен, например, в работе другого исследователя (Donnay 2017), а в качестве примеров подобных исследований можно привести работы (Connelly *et al*. 2016; Donnay *et al*. 2016; Coppock *et al*. 2016) и ряд других.

В данной статье мы применяем методы машинного обучения для эмпирического анализа социально-политической нестабильности. Мы используем данные на страновом уровне с детализацией по годам, и это дает возможность использовать различные статистические источники и формировать большой набор независимых факторов для анализа их влияния на социально-политические процессы.

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 используется 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 такое его 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 точки 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 себя зарекомендовало 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 предсказания для них наиболее интересных и релевантных вакансий (Volkovs *et al.* 2017), или предсказа-

ния 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 Университете штата Нью-Йорк в Бингемтоне, обобщения архива данных *The Statesman's Yearbook,* публикуемых с 1864 г. В базе содержатся данные по более чем 200 странам, а также годовые значения переменных начиная с 1815 г, за исключением периодов 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 мотивированные 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 выдвигались требования, направленные против национальной политики, правительства или органов власти.

К партизанским действиям (*Guerrilla Warfare*, domestic3) относятся любая вооруженная деятельность, диверсии или взрывы, совершаемые 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 направленных 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 демонстрации 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 дестабилизации (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 дестабилизации (*Weighted Conflict Measure*, domestic9) рассчитывается по формуле (4):

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

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 г. (в анализируемой версии по 2015 г. включительно)

Из базы данных государственных переворотов (Marshall M. G., Marshall D. R. 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 гг.

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, классифицированы как факт нестабильности. Для переменной n_terror_attack пороговым значением было выбрано N=50.

В данной статье нас интересуют факторы социально-политической нестабильности для стран, входящих в так называемую Афразийскую зону нестабильности. К данной группе были отнесены 53 страны: Афганистан, Алжир, Армения, Азербайджан, Бахрейн, Бенин, Буркина-Фасо, Центрально-Африканская Республика, Чад, Берег Слоновой Кости, Джибути, Египет, Эритрея, Гамбия, Грузия, Гана, Гвинея, Гвинея-Бисау, Иран, Ирак, Иордания, Казахстан, Кувейт, Киргизия, Ливан, Либерия, Ливия, Мали, Мавритания, Марокко, Нигер, Нигерия, Оман, Пакистан, Палестина, Катар, Саудовская Аравия, Сенегал, Сьерра-Леоне, Сомали, Южный Судан, Судан, Сирийская Арабская Республика, Таджикистан, Идти, Тунис, Турция, Туркменистан, Объединенные Арабские Эмираты, Узбекистан, Йемен, Северный Йемен, Йеменская Народная Республика.

В анализируемой группе стран совместно проживает более 1,2 млрд человек.

В Табл. 2 приведена статистика по 14 зависимым переменным, а также число случаев нестабильности и отсутствия нестабильно-

сти для анализируемых переменных и в странах Афразийской зоны нестабильности:

Табл. 2. Статистика анализируемых переменных нестабильности в странах Афразийской зоны

Переменная	Число наблюдений (страна-год)	Число случаев нестабильности	Число случаев отсутствия нестабильности	Пропущенные данные
cnts_domestic1	2805	272	2533	1157
cnts_domestic2	2805	117	2688	1157
cnts_domestic3	2805	437	2368	1157
cnts_domestic4	2805	310	2495	1157
cnts_domestic5	2805	216	2589	1157
cnts_domestic6	2805	509	2296	1157
cnts_domestic7	2805	508	2297	1157
cnts_domestic8	2805	509	2296	1157
cnts_domestic9	2805	1263	1542	1157
cnts_domestic9_prev	2438	996	1442	1524
cnts_domestic9_next	2454	1012	1442	1508
nkill	2449	780	1669	1513
n_terror_attack	2449	174	2275	1513
coup_detat_failed_coup_detat	3962	215	3747	0

Основной массив данных приходится на временной период с 1950 по 2018 гг. Для нескольких стран присутствуют наблюдения с 1918 г. Столбец «пропущенные данные» показывает пропуски в зависимой переменной для тех периодов (лет), когда в анализируемом массиве данных есть показатели по независимым переменным для интересующей нас страны.

Выбор параметров модели и тренировка моделей

Для оценки (тренировки) модели градиентного бустинга необходимо выбрать набор параметров, определяющих работу алгоритма. Одна из главных проблем, которую необходимо решить при оценке, – 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 получен-

ные 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 улучшения 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 них 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 тестовой 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 20.

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.

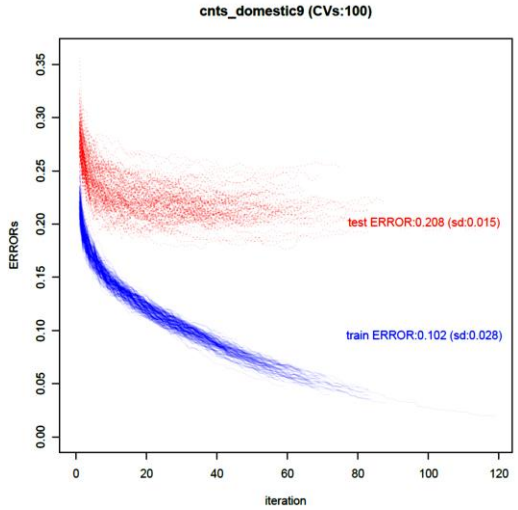

Рис. 1. Динамика ошибок обучающей и тестовой выборки (по всем 300 переменным) для индекса социально-политической дестабилизации (cnts_domestic9) для стран Афразийской зоны нестабильности.

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 Рис. 1 также приводится среднее значение, полученное по всем 100 оценкам. Так, для обучающей выборки (train Error) это значение равно 0,102. То есть на обучающей выборке модель в среднем дает оценку индекса социально-политической дестабилизации, которая ошибается 10,2 % случаев (относительно реального значения индекса социально-политической дестабилизации). 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 дестабилизации (cnts_domestic9).

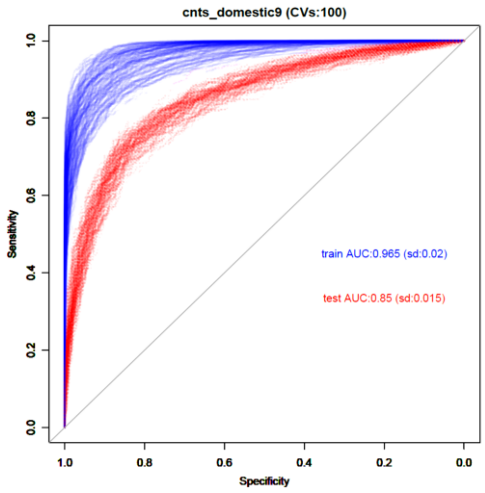

Рис. 2. ROC-кривые для оценок модели (по всем 300 переменным) на обучающих и тестовых выборках для индекса социально-политической дестабилизации (cnts_domestic9) для стран Афразийской зоны нестабильности.

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

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 (Sensitivity) который равен коэффициенту истинно-положительными оценок (TPR). Идеальный классификатор – это классификатор, у которого сочетаются нулевые ложно-положительные ошибки (специфичности = 1) и 100%-ная 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 рис. 2 видно, что модель на обучающей выборке обладает высокими прогностическими возможностями, но и на тестовой выборке ее прогностические способности оказываются не нулевые (AUC = 0,85).

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 на полной базе данных из 300 переменных для стран Афразийской зоны нестабильности

Переменная	Число факторов	Обучающая выборка (train)		Тестовая выборка (test)	
		Error	AUC	Error	AUC
cnts_domestic1	300	0,055 (0,010)	0,907 (0,064)	0,084 (0,009)	0,783 (0,039)
cnts_domestic2	300	0,027 (0,004)	0,869 (0,098)	0,044 (0,008)	0,713 (0,067)
cnts_domestic3	300	0,056 (0,017)	0,975 (0,025)	0,092 (0,010)	0,885 (0,029)
cnts_domestic4	300	0,057 (0,012)	0,927 (0,051)	0,104 (0,012)	0,803 (0,038)
cnts_domestic5	300	0,042 (0,009)	0,934 (0,063)	0,074 (0,009)	0,806 (0,051)
cnts_domestic6	300	0,074 (0,017)	0,950 (0,034)	0,136 (0,012)	0,839 (0,025)
cnts_domestic7	300	0,072 (0,020)	0,943 (0,044)	0,120 (0,011)	0,841 (0,027)
cnts_domestic8	300	0,071 (0,016)	0,952 (0,040)	0,127 (0,013)	0,833 (0,031)
cnts_domestic9	300	0,102 (0,028)	0,965 (0,020)	0,208 (0,015)	0,850 (0,015)
cnts_domestic9_ next	300	0,107 (0,034)	0,960 (0,026)	0,230 (0,017)	0,813 (0,018)
cnts_domestic9_ prev	300	0,102 (0,033)	0,963 (0,025)	0,227 (0,018)	0,817 (0,018)
coup_detat_failed _coup_detat	300	0,024 (0,005)	0,979 (0,015)	0,038 (0,006)	0,945 (0,015)
n_terror_attack	300	0,006 (0,006)	0,993 (0,018)	0,024 (0,006)	0,970 (0,026)
nkill	300	0,056 (0,022)	0,985 (0,013)	0,135 (0,014)	0,914 (0,013)

Примечание: приводятся 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 (количество вершин, в которых используется правило с использованием значения данного параметра).

Каждый из этих факторов может быть использован для оценки относительной важности переменной. Итоговый индекс значимости переменных представлен в формуле (5):

$$importance = (gain + cover + frequency)/3. \quad (5)$$

Размерность коэффициента важности такова, что количественно коэффициент важности (*importance*) на уровне 1 можно интерпретировать как 1%-ный вклад в объясняющую способность модели (через вклад в выигрыш от использования, покрытие данных, частоту использования).

После первого этапа моделирования, на котором мы отранжировали все переменные с использованием показателя общей значимости, из 300 переменных мы отобрали переменные, которые объясняют 75 % от общей значимости.

Так, например, для модели, описывающей переменную «Политические убийства» (cnts_domestic1), распределение общей значимости по числу переменных представлено на Рис. 3.

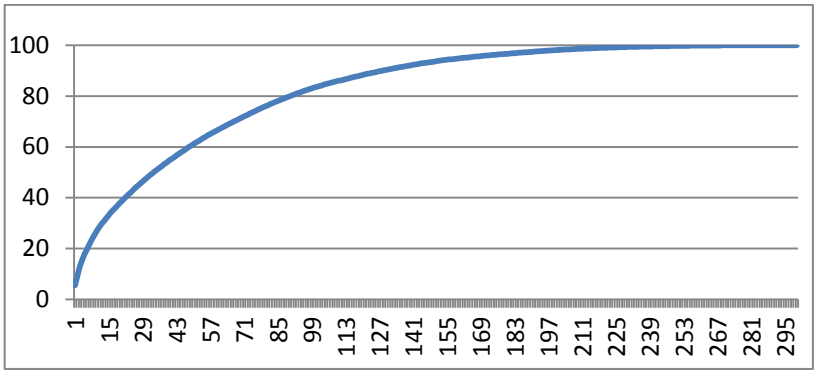

Рис. 3. Распределение общей значимости в зависимости от числа переменных для модели, описывающей переменную «Политические убийства» (cnts_domestic1)

По вертикали отложена общая значимость первых N переменных, а число N отложено по горизонтали. Так, например, первые 33 переменные отвечают за 50 % всей значимости модели. Для упрощения модели и избавления от шума мы в несколько итераций отбрасываем «хвост», на который приходится много незначимых переменных. Так, на первом этапе мы отбираем первые наиболее важные 77 переменных и оцениваем модель с их помощью.

Если сравнить качество модели, то заметим, что оно после первой итерации существенно не изменяется (а для некоторых переменных может даже улучшаться. Например, для cnts_domestic2 или cnts_domestic4 показатели AUC для тестовых выборок повышаются после ограничения числа переменных и сокращения глубины модели).

На втором этапе мы отобрали наиболее значимые факторы и построили модель меньшей глубины (при значении max_depth, равном 4) только на первых N факторах, описывающих 75 % объясняющей способности модели. Например, для модели переменной «Политические убийства» (cnts_domestic1) это 77 переменных, для модели переменной «Политические забастовки» (cnts_domestic2) это 72 переменные, а для модели «Правительственные кризисы» (cnts_domestic4) это первые наиболее важные 74 переменные.

Распределение ROC-кривых, а также динамика ошибок тестовых и обучающих выборок для переменной «Политические убийства» (cnts_domestic1) представлена на Рис. 4:

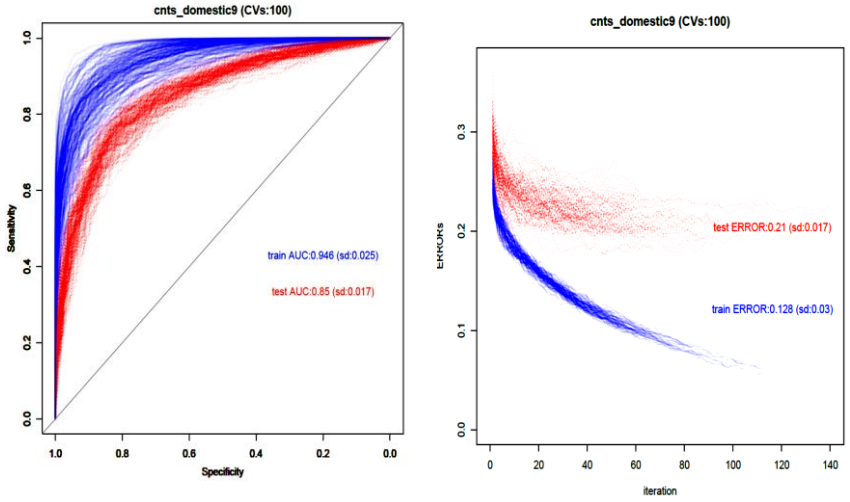

Рис. 4. Оценки качества модели для переменной «Политические убийства» (cnts_domestic1) на обучающих и тестовых выборках после первого шага отбора наиболее значимых факторов для стран Афразийской зоны нестабильности

Табл. 4. Оценки качества моделей на обучающих и тестовых выборках после первого шага отбора наиболее значимых факторов для стран Афразийской зоны нестабильности

Переменная	Число отобранных факторов	Обучающая выборка (train)		Тестовая выборка (test)	
		Error	AUC	Error	AUC
cnts_domestic1	77	0,067 (0,008)	0,876 (0,067)	0,084 (0,009)	0,781 (0,047)
cnts_domestic2	72	0,032 (0,003)	0,838 (0,088)	0,043 (0,008)	0,711 (0,067)
cnts_domestic3	73	0,067 (0,016)	0,962 (0,034)	0,092 (0,010)	0,887 (0,027)

Окончание Табл. 4

Переменная	Число отобранных факторов	Обучающая выборка (train)		Тестовая выборка (test)	
		Error	AUC	Error	AUC
cnts_domestic4	74	0,071 (0,011)	0,912 (0,053)	0,102 (0,011)	0,806 (0,043)
cnts_domestic5	61	0,054 (0,006)	0,907 (0,066)	0,073 (0,010)	0,807 (0,057)
cnts_domestic6	73	0,092 (0,018)	0,928 (0,050)	0,134 (0,012)	0,841 (0,031)
cnts_domestic7	68	0,087 (0,017)	0,924 (0,050)	0,119 (0,012)	0,841 (0,030)
cnts_domestic8	81	0,091 (0,015)	0,926 (0,043)	0,129 (0,014)	0,837 (0,030)
cnts_domestic9	80	0,128 (0,030)	0,946 (0,025)	0,210 (0,017)	0,850 (0,017)
cnts_domestic9_next	81	0,137 (0,034)	0,938 (0,027)	0,231 (0,018)	0,817 (0,019)
cnts_domestic9_prev	74	0,128 (0,032)	0,945 (0,028)	0,225 (0,019)	0,822 (0,019)
coup_detat_failed_ coup_detat	34	0,030 (0,004)	0,975 (0,014)	0,037 (0,006)	0,949 (0,016)
n_terror_attack	56	0,009 (0,008)	0,987 (0,032)	0,023 (0,006)	0,970 (0,034)
nkill	77	0,075 (0,024)	0,975 (0,016)	0,135 (0,014)	0,915 (0,014)

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

Мы еще раз проведем эту процедуру, уже среди отобранных 77 переменных снова выберем подмножество переменных, которое обеспечивает суммарную объясняющую способность в 75 %, и еще уменьшим параметр глубину модели max_depth до 3.

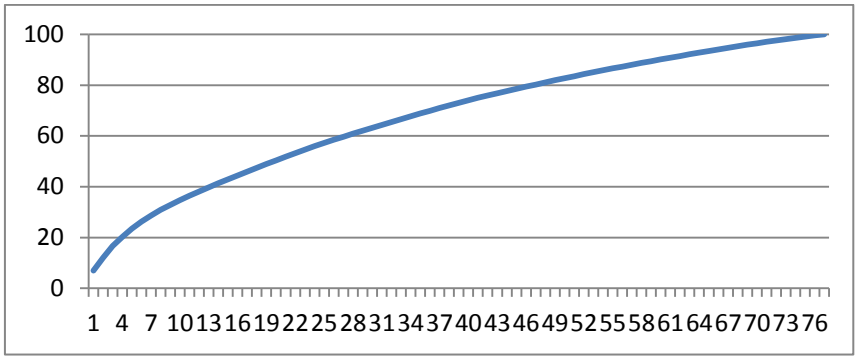

Рис. 5. Распределение общей значимости в зависимости от числа переменных для модели, описывающей переменную «Политические убийства» (cnts_domestic1), после первого шага оптимизации

На следующем этапе для модели, описывающей переменную «Политические убийства» (cnts_domestic1), мы отбираем первые наиболее важные 41 переменную и оцениваем модель с их помощью. Распределение ROC-кривых, а также динамика ошибок тестовых и обучающих выборок после 2-й итерации для переменной «Политические убийства» (cnts_domestic1) представлена на Рис. 6.

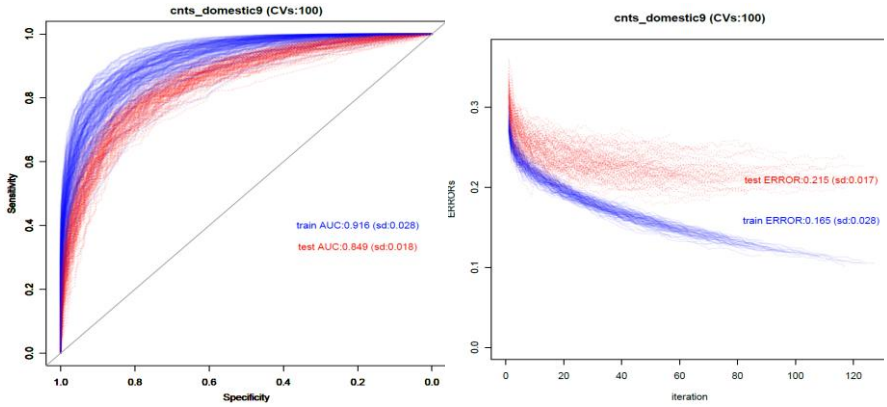

Рис. 6. Оценки качества модели для переменной «Политические убийства» (cnts_domestic1) на обучающих и тестовых выборках после второго шага отбора наиболее значимых факторов для стран Афразийской зоны нестабильности.

После 2-й итерации упрощения заметно упали показатели качества на обучающей выборке, при этом показатели на тестовой выборке остались на сопоставимом уровне, то есть мы избавляемся от излишнего переобучения модели, избыточная сложность которой не добавляет объясняющей способности (см. Рис. 6 и Табл. 5).

Табл. 5. Оценки качества моделей на обучающих и тестовых выборках после второго шага отбора наиболее значимых факторов для стран Афразийской зоны нестабильности

Переменная	Число отобранных факторов	Обучающая выборка (train)		Тестовая выборка (test)	
		Error	AUC	Error	AUC
cnts_domestic1	41	0,076 (0,007)	0,854 (0,072)	0,084 (0,009)	0,780 (0,055)
cnts_domestic2	38	0,037 (0,002)	0,771 (0,088)	0,042 (0,008)	0,691 (0,077)
cnts_domestic3	39	0,079 (0,012)	0,946 (0,040)	0,095 (0,012)	0,884 (0,033)
cnts_domestic4	40	0,089 (0,014)	0,885 (0,057)	0,104 (0,012)	0,804 (0,049)
cnts_domestic5	33	0,066 (0,005)	0,870 (0,058)	0,073 (0,009)	0,799 (0,056)
cnts_domestic6	37	0,117 (0,014)	0,894 (0,051)	0,136 (0,013)	0,836 (0,037)
cnts_domestic7	38	0,097 (0,012)	0,925 (0,029)	0,120 (0,013)	0,853 (0,026)
cnts_domestic8	43	0,109 (0,012)	0,897 (0,038)	0,127 (0,013)	0,834 (0,027)
cnts_domestic9	46	0,165 (0,028)	0,916 (0,028)	0,215 (0,017)	0,849 (0,018)
cnts_domestic9_ next	50	0,167 (0,032)	0,913 (0,028)	0,233 (0,016)	0,816 (0,021)
cnts_domestic9_ prev	42	0,175 (0,024)	0,905 (0,025)	0,230 (0,019)	0,821 (0,019)
coup_detat_failed_ coup_detat	14	0,035 (0,002)	0,964 (0,013)	0,036 (0,006)	0,949 (0,018)
n_terror_attack	27	0,013 (0,008)	0,993 (0,019)	0,024 (0,007)	0,979 (0,020)
nkill	42	0,094 (0,017)	0,960 (0,017)	0,134 (0,015)	0,916 (0,016)

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

Если еще и в 3-й раз применить эту процедуру и выбрать 24 переменные со 2-го шага, объясняющие 75 % для модели зависимой переменной «Политические убийства» (cnts_domestic1), и еще раз снизить глубину модели (max_depth = 2), то заметим снижение в некоторых переменных, однако наибольшее падение по-прежнему наблюдается в показателях обучающей выборки. После данного шага практически не осталось малозначимых переменных.

Распределение ROC-кривых, а также динамика ошибок тестовых и обучающих выборок после 3-й итерации для переменной «Политические убийства» (cnts_domestic1) представлена на Рис. 7.

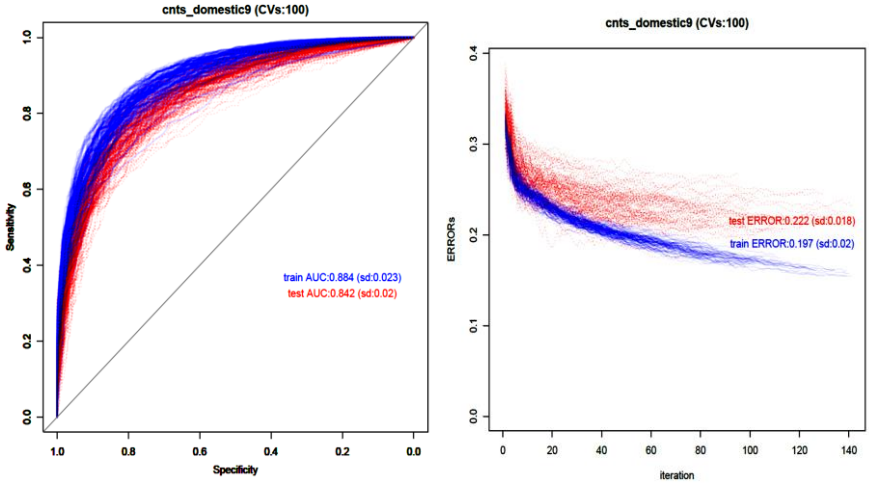

Рис. 7. Оценки качества модели для переменной «Политические убийства» (cnts_domestic1) на обучающих и тестовых выборках после третьего шага отбора наиболее значимых факторов для стран Афразийской зоны нестабильности

После 3-й итерации продолжили падать показатели качества оценки модели для обучающей выборки, а на тестовой выборке они остаются на сопоставимом уровне (см. Рис. 7 и Табл. 6)

Табл. 6. Оценки качества моделей на обучающих и тестовых выборках после третьего шага отбора наиболее значимых факторов для стран Афразийской зоны нестабильности

Переменная	Число отобранных факторов	Обучающая выборка (train)		Тестовая выборка (test)	
		Error	AUC	Error	AUC
cnts_domestic1	24	0,084 (0,004)	0,822 (0,058)	0,086 (0,009)	0,771 (0,050)
cnts_domestic2	21	0,041 (0,002)	0,721 (0,076)	0,041 (0,008)	0,673 (0,069)
cnts_domestic3	23	0,090 (0,009)	0,923 (0,038)	0,098 (0,013)	0,876 (0,035)
cnts_domestic4	23	0,104 (0,008)	0,830 (0,063)	0,105 (0,013)	0,776 (0,071)
cnts_domestic5	19	0,076 (0,003)	0,789 (0,068)	0,076 (0,009)	0,748 (0,067)
cnts_domestic6	21	0,133 (0,012)	0,864 (0,046)	0,139 (0,013)	0,825 (0,041)
cnts_domestic7	22	0,114 (0,013)	0,888 (0,030)	0,127 (0,013)	0,838 (0,025)
cnts_domestic8	23	0,123 (0,013)	0,860 (0,043)	0,131 (0,013)	0,824 (0,033)
cnts_domestic9	28	0,197 (0,020)	0,884 (0,023)	0,222 (0,018)	0,842 (0,020)
cnts_domestic9_next	32	0,219 (0,021)	0,863 (0,023)	0,248 (0,019)	0,808 (0,024)
cnts_domestic9_prev	25	0,206 (0,016)	0,871 (0,017)	0,237 (0,021)	0,816 (0,020)
coup_detat_failed_ coup_detat	5	0,041 (0,002)	0,939 (0,005)	0,041 (0,006)	0,937 (0,015)
n_terror_attack	15	0,026 (0,009)	0,979 (0,045)	0,032 (0,009)	0,963 (0,059)
nkill	25	0,121 (0,016)	0,938 (0,017)	0,143 (0,016)	0,908 (0,017)

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

Проведя в 4-й раз упрощение, мы убедимся, что для глубины max_depth = 1 произошло существенное сокращение показателей

уже и на тестовой выборке, то есть самая простая версия модели может быть адекватно сформулирована на глубине max_depth = 2, допускающей перекрестное использование до двух факторов. Сокращение числа переменных также будет ухудшать показатели тестовой выборки (хотя и не столь существенно, как дальнейшее сокращение сложности моделирования).

Мы остановимся на результатах после 3-го шага упрощения для модели с глубиной (max_depth = 2) и от 5 до 32 наиболее значимых переменных для стран Афразийской макрозоны.

Оценка 100 моделей кросс-валидации после 3 итераций по сокращению незначимых факторов позволяет выделить следующие основные факторы, влияющие на переменную «Политические убийства» (cnts_domestic1):

1. Население в возрасте от 5 до 9 лет (за 5 лет) (X5__9 [10.9]).
2. Долговечность режима (p_durable [7.5]).
3. Количество городского населения, в тыс. человек (e_miurbpop [7.1]).
4. ВВП, в текущей национальной валюте (gdp_current_LCU [6.8]).
5. Финансирование программ со стороны Америки, в долларах (us_foreign_aid [6.1]).

Вторая по важности группа переменных (с коэффициентом importance > 3): население в возрасте от 90 до 94 лет (за 5 лет) (X90__94), индекс регулирования участия (p_parreg), индекс институционализированной автократии (p_autoc), доля голодающих в общей численности населения (%) (share_hungry_population), месяц начала события (MOBEGIN), городское население в возрасте от 80 лет (за 5 лет), в тыс. человек (X80_urban), темпы роста городского населения, в % (urban population growth WB), индекс верховенства права (wdi_rule_of_law), сколько лет текущий глава государства занимает свой пост (bnn_yroff), число военных, на 1000 человек (cnts_size_of_military_population).

Третья группа состоит из 9 факторов (для которых коэффициент важности importance > 1): возраст государственности, независимость (nationality_sovereignty), уровень смертности среди детей, в % (e_peinfmor), свобода академического и культурного самовыражения (v2clacfree), коррупционность судебных решений (v2jucorrdc), плотность населения (cnts_pop_density), темпы роста ВВП, в % (0,01) (e_migdpgro), темпы роста ПИИ, приток

(fdi_inward_percent_gdp), доля населения в возрасте от 15 до 24 лет (share_15__24), население (cnts_pop1).

Данные факторы совместно объясняют 100 % общей значимости модели. Второй группой факторов объясняется 39 %, факторами третьей группы – 22 % общей значимости.

Анализ результатов оценки моделей для переменной «Политические забастовки» (cnts_domestic2) позволяет выделить следующие факторы:

1. Население (cnts_pop1 [14.9]).
2. Индекс избирательной демократии (v2x_polyarchy [13.7]).
3. Индекс институционализированной автократии (p_autoc [9.5]).
4. Общее количество выезжающих за границу студентов, в ед. (total_outbound_tertiary_students [7.4]).
5. Свобода академического и культурного самовыражения (v2clacfree [6.6]).

Во вторую группу входят следующие 6 переменных: плотность населения (cnts_pop_density), численность населения по данным Мэддисона (population_Mad), население (population), индекс либеральной демократии (v2x_libdem), население в возрасте от 55 до 59 лет (за 5 лет) (X55__59), численность населения, в тыс. человек (e_mipopula).

Третья группа состоит из 8 факторов (для которых коэффициент importance > 1): индекс фракционности (cnts_polit01), Onsets of nonviolent campaigns in same region (nvc_dosregt), индекс коррумпированности режима (e_v2xnp_regcorr), ВВП на душу населения по ППС, в постоянных ценах 2011 г. (IS_highest_20), финансирование программ со стороны Америки, в долларах (us_foreign_aid), инфляция, в потребительских ценах (inflation consumer prices), количество безработной молодежи в возрасте 20–29 лет, в тыс. человек (unemployment_youth_15_24_ths), доля занятых в промышленности (employment_industry).

В результате нескольких шагов по выделению значимых переменных перечисленные переменные совместно отвечают за 98 % общей значимости модели. Вторая группа факторов объясняет 27 %, третья группа факторов – 19 %.

Оценка 100 моделей кросс-валидации после 3 итераций по сокращению незначимых факторов позволяет выделить следующие

основные факторы, влияющие на переменную «Партизанские действия» (cnts_domestic3):

1. Финансирование программ со стороны Америки, в долларах (us_foreign_aid [7.2]).

2. ВВП на душу населения по ППС, в постоянных ценах 2011 г. (gdp_per_capita_PPP_Mad [7.1]).

3. Количество городского населения, в тыс. человек (e_miurbpop [7]).

4. Любая государственная дискриминация (dispota4_c [6.1]).

5. Тип смены власти (MAGFAIL [6]).

Вторая по важности группа переменных (с коэффициентом важности importance > 3): население (cnts_pop1), количество безработной молодежи в возрасте 20–29 лет, в тыс. человек (unemployment_youth_15_24_ths), индекс глобализации (index_ globalization), возраст государственности, независимость (natio nality_sovereignty), индекс избирательной демократии (v2x_ polyarchy), свобода академического и культурного самовыражения (v2clacfree), индекс институционализированной автократии (p_autoc), уровень урбанизации, в % (0,01) (e_miurbani), доля грамотных (cnts_percent_literate), долговечность режима (p_durable), фракционность элит (factionalized_elites), уровень смертности среди детей, в % (e_peinfmor).

Третья группа состоит из 6 факторов (для которых коэффициент importance > 1): число военных, на 1000 человек (cnts_size_of_military_population), свобода религии (v2clrelig), беженцы (refugees), доля безработных в молодежи (по данным International Labour Organization [ILO]) (unemployment_youth_total_ ILO), ВВП по ППС, в постоянных долларах США 2011 г. (gdp_PPP), ВВП на душу населения, в ед. (e_migdppc).

Данные факторы совместно объясняют 100 % общей значимости модели. Второй группой факторов объясняется 51 %, факторами третьей группы – 15 % общей значимости.

Тренировка 100 бустинговых моделей позволяет выделить следующие основные факторы, влияющие на зависимую переменную «Правительственные кризисы» (cnts_domestic4):

1. Сколько лет текущий глава государства занимает свой пост (bnn_yroff [8.1]).

2. Доля занятых в промышленности, в процентах (cnts_percent_work_force_in_industry [6.3]).

3. Доля занятых в сфере услуг, в процентах (cnts_percent_ work_force_in_other_activity [6.2]).

4. Индекс институционализированной автократии (p_autoc [6.1]).

5. Тип режима [по Голдстоуну] (p_Goldstone [5.5]).

Вторая по важности группа переменных (с коэффициентом важности importance > 3): год конца события (YREND), свобода академического и культурного самовыражения (v2clacfree), смертность от твердого топлива внутри помещений, (death_rate_ozone), темпы роста ВВП за год (gdp_annual_growth), индекс либеральной демократии (v2x_libdem), финансирование программ со стороны Америки, в долларах (us_foreign_aid), население (cnts_pop1), темпы роста ВВП, в % (0,01) (e_migdpgro), количество лет обучения взрослого населения, в ед. (e_peaveduc), долговечность режима (p_durable), цена на серебро (Silver_price), индекс регулирования участия (p_parreg), количество студентов, всего (students_all).

Третья группа из 5 факторов (для которых коэффициент важности importance > 1): индекс исполнительных ограничений (p_xconst), ВВП на душу населения, в ед. (e_migdppc), число поступивших в вузы, на 1000 человек (cnts_university_enrollment), доля населения в возрасте от 15 до 29 лет (share_15__29), число военных, на 1000 человек (cnts_size_of_military_population).

Последовательно выделяя значимые переменные, мы получили следующий результат: описанные факторы совместно отвечают за 100 % общей значимости модели. Вторая группа факторов отвечает за 54 % общей важности модели, третья группа – за 14 %.

Анализ результатов оценки моделей для переменной «Политические репрессии» (cnts_domestic5) позволяет выделить следующие факторы:

1. Индекс потребительских цен (consumer_price_index [12.7]).

2. Доля занятых в сфере услуг, в % (cnts_percent_work_ force_in_other_activity [12.3]).

3. Индекс цен на товарные продукты питания и напитки (com modity_food_beverage [8.8]).

4. Доля занятых в промышленности, в % (cnts_percent_ work_force_in_industry [8.7]).

5. Цена на серебро (Silver_price [7.4]).

Во вторую группу входят следующие 8 переменных: число поступивших в вузы, на 1000 человек (cnts_university_enrollment), население в возрасте от 70 до 74 лет (за 5 лет) (X70__74), сколько лет текущий глава государства занимает свой пост (bnn_yroff), доля импорта товаров и услуг в ВВП (import_percent_gdp), количество безработной молодежи в возрасте 20–29 лет, в тыс. человек (unemployment_youth_15_24_ths), индекс глобализации (index_ globalization), цена на золото (Gold_price), доля населения в возрасте от 15 до 24 лет (share_15_24).

Третья группа из 6 факторов (для которых коэффициент importance > 1): доля промышленности в ВВП (share_industry), индекс цен на товарные продукты питания (commodity_food), население (cnts_pop1), доля экспорта товаров и услуг в ВВП (export_ percent_gdp), темпы роста ВВП на душу населения за год (gdp_per_ capita_annual_growth), Power distributed by socioeconomic position (v2pepwrses).

В результате нескольких шагов по выделению значимых переменных перечисленные переменные совместно отвечают за 100 % общей значимости модели. Вторая группа факторов объясняет 37 %, третья группа факторов – 13 %.

Анализ результатов оценки моделей для переменной «Массовые беспорядки» (cnts_domestic6) позволяет выделить следующие факторы:

1. Население (cnts_pop1 [11.9]).
2. Число поступивших в вузы, на 1000 человек (cnts_ university_enrollment [6.8]).
3. Численность населения по данным Мэддисона (population_ Mad [6.6]).
4. Население в возрасте от 0 до 4 лет (за 5 лет) (X0__4 [5.9]).
5. Индекс институционализированной автократии (p_autoc [5.6]).

Во вторую группу входят следующие 15 переменных: финансирование программ со стороны Америки, в долларах (us_foreign_aid), цена на золото (Gold_price), цена на нефть марки *Brent* в номинальных ценах (Brent_price_nominal), индекс либеральной демократии (v2x_libdem), количество городов на душу населения (cnts_population_cities_over_per_capita), цена на нефть марки *Brent*, скорректированная на индекс потребительских цен (Brent_price_adjusted_consumer), плотность населения (cnts_pop_

density), темпы роста ВВП за год (gdp annual growth), любая государственная дискриминация (dispota4_c), количество солнечных пятен (sunspot_numbers), ВВП на душу населения по ППС, в постоянных ценах 2011 г. (gdp_per_capita_PPP_Mad), индекс потребительских цен (consumer_price_index), долговечность режима (pol_durable), городское население в возрасте от 75 до 79 лет (за 5 лет), в тыс. человек (X75__79_urban), доля населения в возрасте от 15 до 64 лет (population_ages_15_64).

Третья группа состоит из 1 фактора (для которого коэффициент важности importance > 1): индекс избирательной демократии (v2x_ polyarchy).

В результате нескольких шагов по выделению значимых переменных перечисленные переменные совместно отвечают за 100 % общей значимости модели. Вторая группа факторов объясняет 61 %, третья группа факторов – 2 %.

Тренировка 100 бустинговых моделей позволяет выделить следующие основные факторы, влияющие на зависимую переменную «Государственные перевороты и попытки переворотов» (cnts_domestic7):

1. Число попыток переворотов за предшествующие 5 лет (cou_ tries5 [11.8]).

2. Сколько лет текущий глава государства занимает свой пост (bnn_yroff [8.3]).

3. Индекс регулирования главного исполнительного органа (p_xrreg [8.2]).

4. Население в возрасте от 90 до 94 лет (за 5 лет) (X90__94 [6.5]).

5. Год конца события (YREND [5.8]).

Вторая по важности группа переменных (с коэффициентом importance > 3): доля городского населения в возрасте от 15 до 24 лет (share_urban_15__24), доля безработных среди молодежи (по данным ILO) (unemployment_youth_total_ILO), индекс либеральной демократии (v2x_libdem), экспорт на душу населения, в долларах (cnts_exports_per_capita), ВВП, в текущей национальной валюте (gdp_current_LCU), темпы роста ВВП за год (gdp_annual_growth), доля населения в возрастах от 15 до 24 лет (share_15_24), количество городского населения, в тыс. человек (e_miurbpop), уровень смертности среди детей, в % (e_peinfmor), доля голодающих в об-

щей численности населения (%) (share hungry population), доля валовых внутренних инвестиций в ВВП (gross_fixed_capital_formation_percent_gdp), долговечность режима (pol_durable).

Третья группа из 5 факторов (для которых коэффициент importance > 1): доля занятых в промышленности (employment_industry), темпы роста ВВП, в % (0,01) (e_migdpgro), дельта городского населения (delta_urban_population_UN), смертность от окружающих твердых частиц, (death_rate_indoor_solid_fuels), количество студентов, зачисленных на третий уровень образования (v2petersch).

Последовательно выделяя значимые переменные, мы получили результат, что описанные факторы совместно отвечают за 100 % общей значимости модели. Вторая группа факторов отвечает за 47 % общей важности модели, третья группа – за 13 %.

Анализ результатов оценки моделей для переменной «Антиправительственные демонстрации» (cnts_domestic8) позволяет выделить следующие факторы:

1. Число поступивших в вузы, на 1000 человек (cnts_university_enrollment [10.8]).

2. Индекс потребительских цен (consumer_price_index [7.1]).

3. Население в возрасте от 5 до 9 лет (за 5 лет) (X5__9 [6.7]).

4. Количество городов на душу населения (cnts_population_cities_over_per_capita [6.1]).

5. Темпы роста ВВП за год (gdp_annual_growth [5.6]).

Во вторую группу входят следующие 14 переменных: население (cnts_pop1), плотность населения (cnts_pop_density), фракционность элит (factionalized_elites), долговечность режима (pol_durable), численность населения по данным Мэддисона (population Mad), финансирование программ со стороны Америки, в долларах (us_foreign_aid), индекс институционализированной автократии (p_autoc), младенческая смертность (mortality_rate), население в возрасте от 75 до 79 лет (за 5 лет) (X75__79), городское население в возрасте от 45 до 49 лет (за 5 лет), в тыс. человек (X45__49_urban), индекс либеральной демократии (v2x_libdem), беженцы (refugees), свобода академического и культурного самовыражения (v2clacfree), население в возрасте от 60 до 64 лет (за 5 лет) (X60__64).

Третья группа из 4 факторов (для которых коэффициент importance > 1): цена на нефть марки *Brent* в номинальных ценах

(Brent_price_nominal), уровень урбанизации, в % (0,01) (e mi urbani), количество солнечных пятен (sunspot_numbers), индекс избирательной демократии (v2x_polyarchy).

В результате нескольких шагов по выделению значимых переменных перечисленные переменные совместно отвечают за 100 % общей значимости модели. Вторая группа факторов объясняет 56 %, третья группа факторов – 8 %.

Анализ результатов оценки моделей для переменной «Агрегированный индекс социально-политической дестабилизации» (cnts_do mestic9) позволяет выделить следующие факторы:

1. Население (cnts_pop1 [10.9]).

2. Число попыток переворотов за предшествующие 5 лет (cou_ tries5 [6.1]).

3. Долговечность режима (p_durable [5.5]).

4. Сколько лет текущий глава государства занимает свой пост (bnn_yroff [5.1]).

5. Финансирование программ со стороны Америки, в долларах (us_foreign_aid [4.6]).

Во вторую группу входят следующие 10 переменных: индекс цен на товарные продукты питания (commodity_food), месяц начала события (MOBEGIN), доля грамотных (cnts_percent_literate), любая государственная дискриминация (dispota4_c), индекс либеральной демократии (v2x_libdem), доля занятых в промышленности (employment_industry), коррупционность судебных решений (v2jucorrdc), количество городского населения, в тыс. человек (e_miurbpop), население в возрасте от 0 до 4 лет (за 5 лет) (X0__4), доля населения в возрасте от 15 до 29 лет (share_15__29).

Третья группа из 10 факторов (для которых коэффициент importance > 1): суммарный индекс свобод (fh_status_sum), количество городов на душу населения (cnts_population_cities_over_per_ capita), темпы роста городского населения, в % (urban_population_ growth_WB), индекс институционализированной автократии (p_au toc), количество солнечных пятен (sunspot_numbers), количество студентов, всего (students_all), доля городского населения в возрасте от 15 до 24 лет (share_urban_15__24), плотность населения (cnts_pop_density), количество безработной молодежи в возрасте 20–29 лет, в тыс. человек (unemployment_youth_15_24_ths), темпы роста ВВП, в % (0,01) (e_migdpgro).

В результате нескольких шагов по выделению значимых переменных перечисленные переменные совместно отвечают за 95 % общей значимости модели. Вторая группа факторов объясняет 37 %, третья группа факторов – 26 %.

Тренировка 100 бустинговых моделей позволяет выделить следующие основные факторы, влияющие на зависимую переменную «Агрегированный индекс социально-политической дестабилизации (прогноз на будущий год)» (cnts_domestic9_next):

1. Население в возрасте от 0 до 4 лет (за 5 лет) (X0__4 [7.2]).
2. Индекс регулирования участия (p_parreg [5.9]).
3. Любая государственная дискриминация (dispota4_c [4.7]).
4. Население в возрасте от 55 до 59 лет (за 5 лет) (X55__59 [4.5]).
5. Индекс либеральной демократии (v2x_libdem [4.2]).

Вторая по важности группа переменных (с коэффициентом importance > 3): население в возрасте старше 80 лет (за 5 лет) (X80), количество городского населения, в тыс. человек (e_miurbpop), свобода академического и культурного самовыражения (v2clacfree), доля занятых в сфере услуг, в процентах (cnts_percent_work_force_in_other_activity), население в возрасте от 5 до 9 лет (за 5 лет) (X5__9), индекс глобализации (index_globalization), доля занятых в промышленности (employment_industry), доля грамотных (cnts_percent_literate).

Третья группа из 12 факторов (для которых коэффициент importance > 1): уровень смертности среди детей, в % (e_peinfmor), численность населения по данным Мэддисона (population_Mad), младенческая смертность (mortality_rate), число попыток переворотов за предшествующие 5 лет (cou_tries5), месяц начала события (MOBEGIN), темпы роста городского населения, в % (urban population_growth_WB), темпы роста ПИИ, отток (fdi_outward_percent_gdp), доля городского населения в возрасте от 20 до 29 лет (share_urban_20__29), индекс верховенства права (wdi_rule_of_law), долговечность режима (p_durable), ВВП на душу населения по ППС, в постоянных ценах 2011 г. (gdp_per_capita_PPP_Mad), инфляция, в потребительских ценах (inflation_consumer_prices).

Последовательно выделяя значимые переменные, мы получили следующий результат: описанные факторы совместно отвечают за 87 % общей значимости модели. Вторая группа факторов отвечает за 28 % общей важности модели, третья группа – за 32 %.

Анализ результатов оценки моделей для переменной «Агрегированный индекс социально-политической дестабилизации (с лагом 1 год)» (cnts_domestic9_prev) позволяет выделить следующие факторы:

1. Число попыток переворотов за предшествующие 5 лет (cou_ tries5 [8.9]).
2. Население в возрасте от 0 до 4 лет (за 5 лет) (X0__4 [7.3]).
3. Долговечность режима (p_durable [7.2]).
4. Население (cnts_pop1 [6.4]).
5. Индекс либеральной демократии (v2x_libdem [5.9]).

Во вторую группу входят следующие 11 переменных: число военных, на 1000 человек (cnts_size_of_military_population), количество городского населения, в тыс. человек (e_miurbpop), индекс потребительских цен (consumer_price_index), население в возрасте от 75 до 79 лет (за 5 лет) (X75__79), доля населения в возрасте от 15 до 29 лет (share_15__29), любая государственная дискриминация (dispota4_c), экспорт на душу населения, в долларах (cnts_exports_per_capita), финансирование программ со стороны Америки, в долларах (us_foreign_aid), производство товаров и услуг (wdi_manuf_mi), цена на серебро (Silver_price), доля городского населения в возрасте от 15 до 24 лет (share_urban_15__24).

Третья группа из 9 факторов (для которых коэффициент importance > 1): доля импорта товаров и услуг в ВВП (import_ percent_gdp), темпы роста населения, в % (cnts_pop_ growth), темпы роста городского населения, в % (urban_population_ growth_WB), дельта доли городского населения (delta_share_ urban_population_UN), количество лет обучения взрослого населения, в ед. (e peaveduc), доля валовых внутренних инвестиций в ВВП (gross_fixed_capital_formation_percent_gdp), население в возрасте от 5 до 9 лет (за 5 лет) (X5__9), сколько лет текущий глава государства занимает свой пост (bnn_yroff), доля населения в возрастах от 15 до 24 лет (share_15_24).

В результате нескольких шагов по выделению значимых переменных перечисленные переменные совместно отвечают за 100 % общей значимости модели. Вторая группа факторов объясняет 41 %, третья группа факторов – 23 %.

Тренировка 100 бустинговых моделей позволяет выделить следующие основные факторы, влияющие на зависимую переменную

«Государственные перевороты и попытки переворотов» (coup_detat_failed_coup_detat):

1. Число попыток переворотов за предшествующие 5 лет (cou_ tries5 [59.4]).

2. Индикатор типа смены режима (POLITYX [15.4]).

3. Сколько лет текущий глава государства занимает свой пост (bnn_yroff [14]).

4. Год начала события (YRBEGIN [11]).

5. Долговечность режима (p_durable [0.2]).

Последовательно выделяя значимые переменные, мы получили следующий результат: описанные факторы совместно отвечают за 100 % общей значимости модели.

Тренировка 100 бустинговых моделей позволяет выделить следующие основные факторы, влияющие на зависимую переменную «Количество террористических атак» (n_terror_attack):

1. Население (population [17.6]).

2. Финансирование программ со стороны Америки, в долларах (us_foreign_aid [11.4]).

3. Индекс борьбы с коррупцией (wdi_control_of_corruption [10.4]).

4. ВВП, в текущей национальной валюте (gdp_current_LCU [6.5]).

5. Комбинированная оценка Polity IV (p_polity_2_2 [6]).

Вторая по важности группа переменных (с коэффициентом importance > 3): возраст государственности, независимость (nationa lity_sovereignty), индекс институционализированной автократии (p_autoc), коэффициент смертности (mort_rate), доля занятых в сфере услуг (employment_services), год начала события (YRBEGIN), индекс регулирования участия (p_parreg), плотность населения (cnts pop density), доля безработных в молодежи (по данным ILO) (unemployment_youth_total_ILO), месяц конца события (MOEND), численность населения по данным Мэддисона (population_Mad).

Последовательно выделяя значимые переменные, мы получили результат: описанные факторы совместно отвечают за 100 % общей значимости модели. Вторая группа факторов отвечает за 48 % общей важности модели.

Анализ результатов оценки моделей для переменной «Количество убитых» (nkill) позволяет выделить следующие факторы:

1. Долговечность режима (p_durable [8.4]).

2. Население (population [8.1]).

3. Freedom of religion (v2clrelig [5.9]).

4. Население в возрасте от 0 до 4 лет (за 5 лет) (X0__4 [5.2]).

5. Индекс цен на товарные продукты питания и напитки (com modity_food_beverage [5]).

Во вторую группу входят следующие 13 переменных: доля женщин среди городского населения 20–29 лет (share_urban_20__ 29_f), ВВП по ППС, в постоянных долларах США 2011 г. (gdp_ PPP), коэффициент смертности (mort_rate), плотность населения (cnts_pop_density), количество безработной молодежи в возрасте 20–29 лет, в тыс. (unemployment_youth_15_24_ths), доля безработных в молодежи (по данным ILO) (unemployment_youth_total_ILO), возраст государственности, независимость (nationality_sovereignty), индекс институционализированной автократии (p_autoc), возраст государственности, появление государства или квазигосударственного образования (nationality_appearance), доля населения в возрастах от 15 до 64 лет (population_ages_15_64), доля занятых в сельском хозяйстве (employment_agriculture), доля голодающих в общей численности населения (%) (share_hungry_population), беженцы (refugees).

Третья группа из 7 факторов (для которых коэффициент importance > 1): политическая значимость этнической принадлежности элит (elceleth_c), индекс регулирования участия (p_parreg), ВВП, в текущей национальной валюте (gdp_current_LCU), индекс гражданских свобод (fh_cl), количество солнечных пятен (sunspot_ numbers), темпы роста ПИИ, приток (fdi_inward_percent_gdp), финансирование программ со стороны Америки, в долларах (us_foreign_aid).

В результате нескольких шагов по выделению значимых переменных перечисленные переменные совместно отвечают за 100 % общей значимости модели. Вторая группа факторов объясняет 49 %, третья группа факторов – 19 %.

Три итерации по сокращению незначимых факторов и тренировка 1400 бустинговых моделей кросс-валидации (по 100 по каждой из 14 анализируемых переменной) позволяет выделить следующие основные независимые факторы, влияющие на все переменные социально-политической нестабильности:

1. Число попыток переворотов за предшествующие 5 лет (cou_ tries5 [6.2]).

2. Население (cnts_pop1 [4.3]).

3. Финансирование программ со стороны Америки, в долл. (us_foreign_aid [3.5]).

4. Индекс институционализированной автократии (p_autoc [3.3]).

5. Сколько лет текущий глава государства занимает свой пост (bnn_yroff [3.2]).

6. Долговечность режима (p_durable [3]).

7. Индекс либеральной демократии (v2x_libdem [2.6]).

8. Население (population [2.5]).

9. Количество городского населения, в тыс. (e_miurbpop [2.3]).

10. Плотность населения (cnts_pop_density [2.1]).

В следующую по важности группу входят следующие 20 факторов: долговечность режима (p_durable), индекс либеральной демократии (v2x_libdem), население (population), количество городского населения, в тыс. человек (e_miurbpop), плотность населения (cnts_pop_density), население в возрасте от 0 до 4 лет (за 5 лет) (X0__4), число поступивших в вузы, на 1000 человек (cnts_university_enrollment), индекс потребительских цен (consumer_price_index), свобода академического и культурного самовыражения (v2clacfree), население в возрасте от 5 до 9 лет (за 5 лет) (X5__9), численность населения по данным Мэддисона (population_Mad), индекс регулирования участия (p_parreg), ВВП, в текущей национальной валюте (gdp_current_LCU), индекс избирательной демократии (v2x_polyarchy), любая государственная дискриминация (dispota4_c), темпы роста ВВП за год (gdp_annual_growth), доля занятых в сфере услуг, в % (cnts_percent_work_force_in_other_activity), количество безработной молодежи в возрасте 20–29 лет, в тыс. человек (unemployment_youth_15_24_ths), индикатор типа смены режима (POLITYX), год начала события (YRBEGIN).

Данные факторы совместно объясняют 57 % общей значимости модели. Второй группой факторов объясняется 36 % общей значимости.

Для анализа региональных особенностей и сопоставления с Афразийской зоной нестабильности мы провели аналогичный анализ и для подвыбороки стран, относящихся к другим регионам: Ла-

тинской Америке, Восточной Европе, странам Африки южнее Сахеля и подгруппе наиболее развитых западных стран.

В отличие от стран Афразийской макрозоны для группы стран Латинской Америки дополнительно среди наиболее важных факторов, определяющих различные формы социально-политической нестабильности, появляется фактор рыночной конъюнктуры: цена на нефть марки *Brent* в номинальных ценах и такие характеристики, как индекс политической фракционности и доля занятых в сфере услуг. Среди наиболее значимых отсутствуют такие факторы, как индекс глобализации.

Для группы стран Восточной Европы в отличие от стран Афразийской макрозоны в качестве наиболее важных факторов добавляются доля занятых в промышленности, экспорт на душу населения, а также степень коррумпированности режима. Как и для стран Латинской Америки, важным фактором выступает цена на нефть марки *Brent* в номинальных ценах.

Заключение и обсуждение результатов

Мы использовали процедуру последовательного выделения набора значимых факторов, чтобы исключить влияние случайных факторов (объясняющих «случайный шум» с помощью другого «случайного шума»). Последовательно упрощая модель и сокращая число факторов, мы сформулировали класс моделей, содержащих 10–30 наиболее значимых факторов, и оценили их значимость с помощью их вклада в объясняющую способность модели.

Используя методы градиентного бустинга, мы отобрали набор наиболее важных переменных для мониторинга, необходимых для оценки политической нестабильности.

В текущей версии мы использовали модели, в которых анализируется лишь один из аспектов нестабильности (измеряемый с помощью индексов нестабильности *CNTS*).

В данной работе мы использовали модель «common pool» для анализа наиболее важных факторов, влияющих на нестабильность, построение моделей, учитывающих временную структуру факторов, позволит в дальнейшем уточнить набор наиболее важных факторов для оценки будущих социально-политических рисков.

Библиография

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 г. *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:* ежегодник / Отв. ред. Л. Е. Гринин, А. В. Коротаев, Л. М. Исаев, А. Р. Шишкина. Волгоград: Учитель. С. 57–83.

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

Chen T., Guestrin C. 2016. Xgboost: A scalable tree boosting system. *Proceedings of the 22nd acmsigkdd international conference on knowledge discovery and data mining.* ACM. Pp. 785–794.

Connelly R., Playford C. J., Gayle V., Dibben C. 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. Vol. 8(1).

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

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: Science Application International Corporation. Failed and Fragile States. URL: http://www4.carleton.ca/cifp/.

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

Korotaev A., Shulgin S., Zinkina J. 2017. Анализ страновых рисков с использованием демографических и социально-экономических данных. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2944064.

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

Volkovs M., Wei Yu G., Poutanen T. 2017. Content-based Neighbor Models for Cold Start in Recommender Systems. *Proceedings of the Recommender Systems Challenge*. Como, Italy. P. 7.

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

Mironov V., A. Guschin. 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/.

Przeworski A., Alvarez M. E.; Cheibub J. A., Limongi F. 2000. *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. URL: arXiv preprint arXiv:1609.02728.

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. URL: https://www.start.umd.edu/gtd/.