УДК 004.8 DOI: 10.26102/2310-6018/2021.33.2.030

# Использование методов машинного обучения для прогнозирования раскрываемости преступлений на основе документов первичного учета

**Д.Ю. Булгаков**

*Федеральное казенное учреждение «Главный информационно-аналитический центр Министерства внутренних дел Российской Федерации», Москва, Российская Федерация*

***Резюме:*** Раскрываемость преступлений является одним из важных показателей деятельности органов внутренних дел. Несмотря на совершенствование методов противодействия преступности и документирования преступной деятельности, раскрываемость преступлений в Российской Федерации остается на уровне 51 %-56 %. В статье описан метод построения математической модели – цифрового двойника зарегистрированного преступления. В качестве исходных данных для построения модели использован массив сведений – документов первичного учета, в отношении 341 тысячи преступлений, совершенных на территории Приморского края за 11 лет – с 2010 по 2020 годы. Модель позволяет: с достоверностью 88 %, на основании формализованной первичной информации, содержащейся в статистических карточках Форма № 1 «На выявленное преступление», строить прогноз о том, будет преступление раскрыто или нет; проводить ревизию нераскрытых преступлений прошлых лет в целях определения преступлений, имеющих высокую вероятность раскрытия; выявлять признаки в статистических карточках, в наибольшей степени влияющие на прогноз раскрываемости преступлений. Модель основана на использовании алгоритмов машинного обучения «градиентный бустинг над решающими деревьями», реализованных в открытой библиотеке искусственного интеллекта CatBoost от компании Яндекс. Точность модели подтверждается составлением и проверкой прогноза раскрываемости в январе-июне 2021 года, 16408 преступлений, совершенных на территории Приморского края.

***Ключевые слова:*** цифровой двойник, прогностическая модель, преступность, статистические карточки, машинное обучение, искусственный интеллект, CatBoost, градиентный бустинг, решающие деревья, значимость признаков.

***Для цитирования:*** Булгаков Д.Ю., Использование методов машинного обучения в прогнозировании раскрываемости преступлений на основе документов первичного учета. *Моделирование, оптимизация и информационные технологии.* 2021;9(2). Доступно по: https://moitvivt.ru/ru/journal/pdf?id=1010 DOI: 10.26102/2310-6018/2021.33.2.030

# Using machine learning methods to predict the detection of crimes based on primary accounting documents

**D.Y. Bulgakov**

*Federal government institution «The Main Informational Analytic Centre of the Ministry of Internal Affairs of the Russian Federation», Moscow, Russian Federation*

***Abstract:*** The result of crime-solving is one of the crucial indicators of the law enforcement communities activities. Despite the improvement of crime investigation methods, the success rate of crime detection in the Russian Federation remains at a level of 51%–56%. The article describes a technique for constructing a mathematical model – a digital double of a registered crime. As the initial data for the model development, an array of information - primary accounting documents about 341 thousand crimes, committed on the territory of the Primorsky Krai over 11 years from 2010 to 2020.

The model allows you: with 88% confidence, based on the formalized primary information contained in the primary accounting documents – statistical cards Form No. 1 On the detected crime, to make a forecast about whether the crime will be solved or not; to audit unsolved crimes of previous years to determine the crimes that have a high probability of detection; to identify the features in the statistical cards that most affect the forecast of the detection of crimes. The model is based on machine learning algorithms “gradient boosting over decision trees”, implemented in the open library of artificial intelligence CatBoost from Yandex. The model accuracy is confirmed by the preparation and verification of the investigation of the crime's result forecast in January–June 2021 for 16408 guilty acts committed on the territory of the Primorsky Krai.

***Keywords:*** digital double, predictive model, crime, statistical cards, machine learning, artificial intelligence, CatBoost, gradient boosting, decision trees, feature importance.

***For citation:*** Bulgakov D.Y. Using machine learning methods to predict the detection of crimes based on primary accounting documents *Modeling, Optimization and Information Technology*. 2021;9(2). Available from: https://moitvivt.ru/ru/journal/pdf?id=1010 DOI: 10.26102/2310-6018/2021.33.2.030 (In Russ).

## Введение

В России, на протяжении последних 5 лет [1] всеми правоохранительными органами ежегодно регистрируется в среднем 2,05 млн преступлений (2016 год – 2 160 063, 2017 – 2 058 476, 2018 – 1 991 532, 2019 – 2 024 337, 2020 – 2 044 221). Подавляющее большинство из этих преступлений – 1,92 млн (93,6%) выявляется сотрудниками органов внутренних дел (далее – ОВД).

В зависимости от подследственности, зарегистрированные преступлениям расследуются следователями и дознавателями шести правоохранительных органов: СК России, МВД России, ФСБ России, ФТС России, ФССП России, МЧС России. При этом производство по подавляющему числу (89 %) зарегистрированных преступлений осуществляется следователями и дознавателями ОВД.

Успешность расследования преступлений, которая отражается в статистической отчетности как *раскрываемость преступлений* и представляет собой отношение числа *раскрытых* к сумме *раскрытых* и *нераскрытых* в отчетном году преступлений, на протяжении последних 5 лет остается в пределах 51 %-56 %.

В 2016 году [1] раскрываемость составила 54,75 %, в 2017 – 55,76 %, в 2018 – 55,84 %, в 2019 – 53,49 %, в 2020 – 51,71 %. Таким образом, с 2018 по 2020 годы раскрываемость преступлений снизилась с 55,8 % до 51,7 % и достигла исторического минимума за последние 13 лет (2008-2020 годы).

В данной статье используются следующие понятия:

- *Раскрытые преступления* – преступления, раскрытые в процессуальном смысле [2], что соответствует этому понятию с точки зрения единого учета преступлений [3] –преступления, по которым лица, их совершившие, установлены и уголовные дела о данных преступлениях направлены в суд с обвинительным заключением/актом или прекращены по нереабилитирующим основаниям;
- *Нераскрытые преступления* – преступления, независимо от времени совершения и регистрации, следствие по которым в отчетном периоде впервые приостановлено за розыском подозреваемого или обвиняемого (п.2 ч.1 ст.208 УПК РФ); неустановлением лица, подлежащего привлечению в качестве обвиняемого (п.1 ч.1 ст.208 УПК РФ); либо в случае когда место нахождения подозреваемого или обвиняемого известно, однако реальная возможность его участия в уголовном деле отсутствует (п.3 ч.1 ст.208 УПК РФ). Данное понятие регламентировано п.2.11

Положения о едином порядке регистрации уголовных дел и учета преступлений, утвержденного совместным приказом «О едином учете преступлений» [3].

На протяжении последних 5 лет в России ежегодно остаются нераскрытыми от 850 тыс. до 985 тыс. преступлений (нераскрыто преступлений [1]: в 2016 году – 983 355, в 2017 – 886 786, в 2018 – 860 408, в 2019 – 915 204, в 2020 – 963 752).

Нераскрытые по итогам отчетного года преступления, то есть преступления, производство по которым в отчетном периоде впервые приостановлено по пп. 1-3 ч. 1 ст. 208 УПК РФ, ставятся на отдельный учет и составляют отдельную категорию преступлений – *преступления прошлых лет* (далее – ППЛ).

Работа по расследованию ППЛ в последующих годах продолжается, но вероятность того, что нераскрытое ППЛ будет в последующем раскрыто, невелика. Так, если из вновь зарегистрированных преступлений раскрывается примерно каждое второе преступление, то из числа ППЛ в последующем раскрывается лишь каждое двадцатое преступление.

Ежегодно раскрывается от 43 тыс. до 54 тыс. ППЛ (раскрыто ППЛ [1]: в 2016 году – 54 328, в 2017 – 51 196, в 2018 – 48 588, в 2019 – 43 626, в 2020 – 46 378.). ППЛ, по которым истекли сроки давности, снимаются с учета. По состоянию на январь 2021 года (раздел 7 отчета 4-ЕГС за январь 2021 года [1]) на учете находилось 6,44 млн нераскрытых ППЛ, уголовные дела о 5,83 млн (90,5%) из которых находились или находятся в производстве сотрудников ОВД.

Целью исследования являлось построение математической модели – *цифрового двойника* зарегистрированного преступления – прогнозирующей вероятность раскрытия преступлений, основанной на выявлении закономерностей в сведениях, имеющихся лишь на момент регистрации преступления.

Созданная модель позволила бы:

1. На момент регистрации определять преступления, имеющие сильную и слабую перспективу на раскрытие, чтобы планировать работу по расследованию данных преступлений.
2. Периодически проводить ревизию нераскрытых преступлений, в том числе ППЛ, в целях выявления преступлений, имеющих сильную перспективу на раскрытие. Это позволит целенаправленно уделять больше внимания, в том числе повторно, данной категории преступлений.
3. Выявлять районы и регионы ОВД, имеющие схожие исходные данные, но демонстрирующие разные показатели в раскрываемости преступлений. Это позволит более детально анализировать оперативную обстановку и результаты противодействия преступности на соответствующей территории, в том числе в отношении конкретных категорий преступлений.
4. Выявить признаки в сведениях о зарегистрированных преступлениях, оказывающих наибольшее влияние на прогноз раскрываемости в целях анализа причин и условий, влияющих на раскрываемость преступлений.

## Материалы и методы

Перечень сведений, характеризующих различные этапы расследования преступлении и используемых для формирования статистической отчетности о преступности и результатах борьбы с ней, регламентирован совместным приказом «О едином учете преступлений» [3]. В соответствии с данным приказом следователи и дознаватели на различных этапах расследования уголовного дела заполняют формализованные *документы первичного учета* – статистические карточки. В качестве исходных данных для построения модели взяты сведения, содержащиеся в

статистических карточках Формы № 1 «Статистическая карточка на выявленное преступление» и Форма № 1.1 «Статистическая карточка о результатах расследования преступления».

Широкое распространение для анализа структурированных данных и автоматизированного построения математических моделей на их основе получили методы машинного обучения, ввиду возможности автоматизации обработки большого количества информации и более высокой точности в сравнении с традиционными линейными моделями или статистическими методами.

Часто, для анализа сведений о преступности используются методы, основанные на нейронных сетях [4-7]. При этом, использование таких методов накладывает определенные ограничения на исходные данные – как правило, это должны быть хорошо структурированные, очищенные данные, с минимальным количеством пропусков. Особый подход требуется для обработки категориальных признаков – то есть признаков, основанных на словарных значениях. Вызывают сомнение результаты, получаемые при обучении нейронной сети на наборах данных – *datacetmax*, небольшого размера, содержащих лишь сотни записей.

В некоторых технологиях искусственного интеллекта, например в задачах компьютерного зрения, к которым, относится задача распознавания лиц, нейронным сетям нет равных [8, 9].

В то же время, существуют алгоритмы машинного обучения, которые при обработке структурированной информации демонстрируют точность, сопоставимую с точностью нейронных сетей, но при этом для своего обучения требуют значительно меньше вычислительных ресурсов и способны обучаться без использования GPU на датасетах меньшего размера.

Общим подходом построения более точных моделей является применение ансамбля моделей. Исходя из целей решаемой задачи был выбран алгоритм градиентного бустинга над решающими деревьями. Этот алгоритм позволяет строить аддитивную функцию в виде суммы решающих деревьев итерационно. Идея данного метода заключается в минимизации среднеквадратической ошибки при обучении. Данный подход позволяет расширить круг решаемых этим алгоритмом задач, а также, зачастую, получить выигрыш в точности предсказания в сравнении с алгоритмами, не использующими ансамблирование.

Использование категориальных признаков в решающих деревьях обуславливает построение $n$-арных деревьев, то есть таких деревьев, где из каждой вершины могут выходить до $n$ ребер. Если признак вещественный или бинарный, то все еще можно использовать простое условие с порогом $t$, поэтому интерес представляет именно случай категориального признака. В случае категориального признака разбиение вершины происходит по тому признаку, для которого значение критерия ошибки будет минимальным.

Для построения прогностической модели был выбран фреймворк CatBoost [10], реализующий метод градиентного бустинга над решающими деревьями. CatBoost разрабатывается компанией Яндекс, представлен общественности в качестве проекта с открытым исходным кодом в 2017 году под свободной лицензией Apache License 2.0.

CatBoost предназначен для работы с датасетами, содержащими большое количество категориальных признаков и имеет встроенные функции преобразования таких данных, отсюда и происходит его название – «категориальный бустинг». При этом CatBoost, в отличии от некоторых других фреймворков, для обработки категориальных признаков не требуются экспертные оценки [11, с. 33].

Для использования CatBoost или обучения моделей, создаваемых на его основе, доступ в Интернет не требуется. Фреймворк устанавливается локально – в виде

библиотеки расширения к установленному интерпретатору языка программирования Python. Размер дистрибутива фреймворка составляет 65 МБ. Поддерживается обучение и использование моделей как на CPU, так и на GPU.

CatBoost написан на языке программирования C++. Модели, созданные на его основе, могут использоваться в языках программирования C, C++, Python, R, Java через подключение соответствующего программного модуля и вызов внешней библиотеки или же через вызов фреймворка в командной строке с соответствующими параметрами.

Как отмечается исследователями, использующими CatBoost [12]: «Особенностью алгоритма является построение симметричных деревьев, возможность работы с категориальными признаками, кроме того, он позволяет обучаться на относительно небольшом количестве неоднородных данных. CatBoost способен решать такие задачи машинного обучения, как регрессия, классификация, мультиклассификация и ранжирование».

Помимо работы с «классическими» непрерывными, категориальным и бинарными признаками, CatBoost «умеет» работать с текстом. На основе текстовых значений CatBoost может формировать значимые для моделей признаки. Таким образом могут решаться задачи поиска близких по смыслу слов и текстов или задачи классификации текстов.

CatBoost достаточно «лояльно» относится к пропускам в данных, когда некоторые из признаков, на которых построена модель, не указаны. Такие случаи часто встречаются в статистической карточке Формы № 1, так как некоторые из её реквизитов необязательны для заполнения или же заполняются в зависимости от наличия соответствующих сведений. Значения в отсутствующих признаках, перед их обработкой моделью, заменяются на определенные значения в соответствии с выбранными правилами – например, создается новое словарное значение для категориальных признаков, используется медианное значение для числовых признаков.

## Обучение модели

Из Формы № 1 для построения модели, по итогам серии экспериментов, были отобраны 83 признака из 160 с ненулевым влиянием на прогноз раскрываемости. Текстовое описание преступления (фабула) и фамилия следователя/дознавателя не использовались.

Из Формы № 1.1 взят только фактический результат расследования преступления (реквизит № 25) – раскрыто преступление или не раскрыто. К этим двум вариантам были сведены 59 возможных значений, которые может принимать данный реквизит (уголовное дело направлено прокурором в суд с обвинительным заключением/актом; дело прекращено по реабилитирующим/нереабилитирующим основаниям; дело приостановлено и т. д.).

После возбуждения уголовного дела – в процессе его расследования, следователь или дознаватель заполняет и другие статистические карточки, в том числе Форму № 1.2 «На преступление, по которому лицо, его совершившее, установлено», а также указывает дополнительные сведения о преступлении в статистической карточке Формы № 1.1 – на момент принятия решения по преступлению. Но эти сведения становятся известны уже в ходе расследования уголовного дела. Кроме того очевидно, что сведения о наличии подозреваемого лица существенно повышают вероятность раскрытия преступления, поэтому для построения модели использованы лишь сведения из Формы № 1 – то, что известно на момент возбуждения уголовного дела.

В качестве исходных данных для построения модели использованы данные информационного центра УМВД России по Приморскому краю – выгрузка из банка

данных статистических карточек, содержащих сведения о 340 929 преступлениях, совершенных на территории Приморского края за 11 лет – с 2010 по 2020 год. Из них – 227 088 статкарточек Формы № 1 на раскрытые преступления и 113 841 статкарточек на нераскрытые. Данный массив был разделен на *обучающий* и *верификационный* (проверочный) датасеты, содержащие сведения о 280 000 и 60 929 преступлениях соответственно.

Построенная модель является моделью *бинарной классификации* – на основе совокупности признаков, известных на момент регистрации преступления, модель, в отношении этого преступления, делает прогноз – преступление будет раскрыто или преступление не будет раскрыто. При этом в виде значения от 0 до 1 оценивается вероятность того, что преступление будет раскрыто. Если модель оценивает вероятность раскрытия преступления от 0 до 0,5, то считаем, что преступление не будет раскрыто. Если модель оценивает вероятность раскрытия от 0,5 до 1, то считаем, что преступление будет раскрыто.

Способ обучения, по которому проводилось обучение модели – это *«обучение с учителем»*. В процессе обучения модели предъявляется совокупность признаков о конкретном преступлении из *обучающего* датасета и сообщается итоговый результат – преступление раскрыто или нераскрыто. В процессе обучения модель должна подобрать условия в решающих деревьях таким образом, чтобы на основе предъявляемых признаков как можно точнее прогнозировать вероятность отнесения преступления к определенному классу – раскрытому или нераскрытому преступлению.

Для проверки точности обучающейся модели, в процессе её обучения используется *верификационный* датасет. В отношении каждого преступления из *верификационного* датасета модель строит прогноз и проверяет – сошелся её прогноз с реальным результатом или нет. После этого происходит переход на следующую итерацию обучения с целью построения решающего дерева с учетом компенсации ошибок, совершенных в условиях предыдущих деревьев. Достоверность прогноза на *верификационном* датасете обычно ниже, чем на *обучающем*.

В процессе обучения модели, после определенной итерации обучения, может возникать эффект «*переобучения модели*» (до тех пор, пока этого не происходит, модель можно считать «*недообученной*»). После данной итерации точность модели на *обучающем* датасете продолжает возрастать, но точность модели на *верификационном* датасете начинает снижаться. Это происходит, потому что модель начинает всё лучше и лучше подстраиваться под *обучающий* датасет – под совокупность признаков, прогноз по которым ей известен заранее, стремясь создать модель, идеально соответствующую обучающему датасету. В тоже время, в *верификационном* датасете совокупность признаков о преступлениях отличается от *обучающего* датасета, и «идеальная подгонка» модели под *обучающий* датасет начинает отрицательно влиять на точность прогноза на незнакомых данных, содержащихся в *верификационном* датасете.

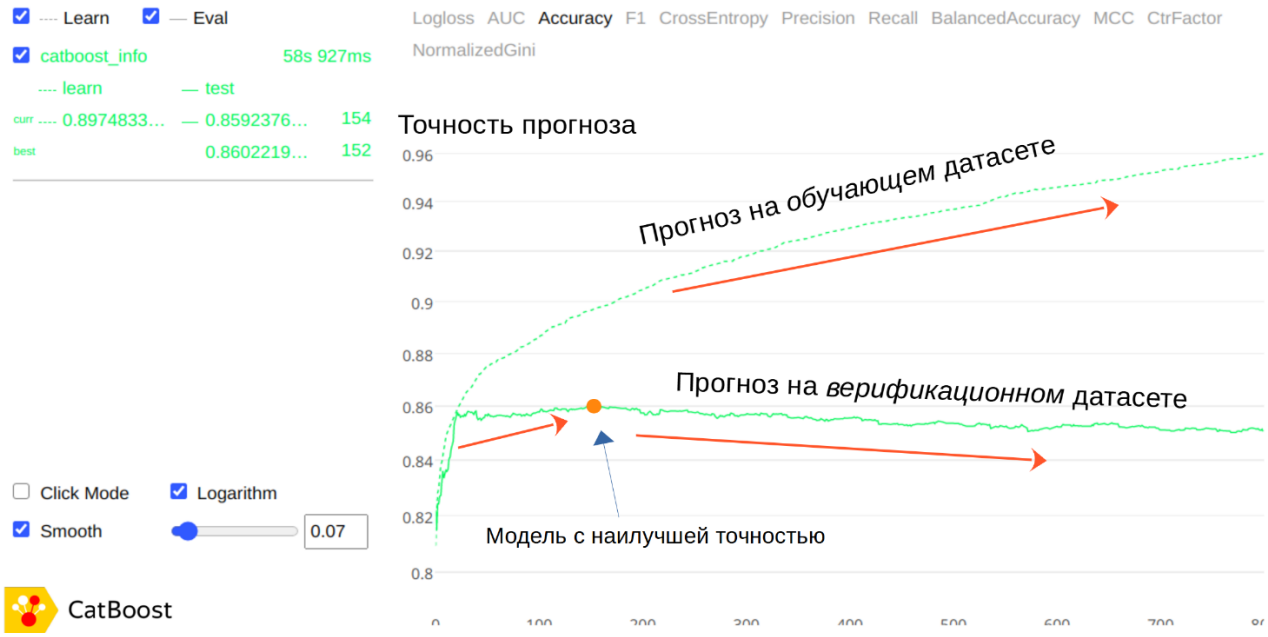


Рисунок 1 – Эффект переобучения модели
Figure 1 – The effect of retraining the model

Процесс обучения модели останавливается, когда становится очевидно, что дальнейшее обучение модели за число заданных деревьев и шага обучения не приводит к возрастанию её точности на *верификационном* датасете. На Рисунке 1 изображен случай, когда в одном из экспериментов максимально достигнутая точность в процессе обучения модели составила 86%. После этой итерации, несмотря на возраставшую точность прогноза на *обучающем* датасете, точность прогноза на *верификационном* датасете постепенно снижалась.

Не всегда имеет смысл дожидаться, когда будет достигнут эффект «переобучения» модели. Модель может становиться всё сложнее и сложнее, число решающих деревьев будет возрастать, но при этом точность модели будет возрастать незначительно.

Итоговое обучение модели, после серии экспериментов, заняло 24 часа на 8 ядрах CPU E5-2630 v3 частотой 2,4 ГГц. GPU не использовался.

Полученная модель построена на 15 000 деревьях. В процессе обучения точкой остановки– лучшей моделью – явилась модель, построенная на итерации 15 000 с минимальным значением логистической функции ошибки на *верификационном* датасете. Увеличение шага обучения приводило к переобучению модели с более низкой метрикой качества.

В качестве программного обеспечения использованы: операционная система Linux CentOS 7.6 x86-64, язык программирования Python 3.7.6, входящий в состав дистрибутива для научных вычислений и машинного обучения Anaconda3 2020.02. Дополнительно установлен фреймворк CatBoost версии 0.23.1.

## Результаты

На верификационной выборке (60 929 преступлений) с соотношением долей раскрытых к нераскрытым преступлениям 2:1, что соответствовало распределению преступлений в обучающей выборке, на построенной модели достигнуты следующие метрики точности [13]: Accuracy = 91,42 %, AUC = 90,77 %, F-мера = 93,58 %.

Если исходную верификационную выборку (60 929 преступлений) привести к сбалансированному виду, где соотношение долей раскрытых к нераскрытым преступлениям будет 1:1, то число преступлений в ней снизится до 40 562, при этом метрики точности построенной модели получатся следующими: Accuracy = 90,77 %, AUC = 90,77 %, F-мера = 90,97 %.

Здесь и далее: *Accuracy* – доля правильных ответов (доля истинных ответов от общего количества ответов), *AUC* – площадь под ROC-кривой, *F-мера* – гармоническое среднее точности и полноты.

Скорость работы модели на 8 ядрах CPU E5-2630 v3 с частотой 2,4 ГГц составляет 1 секунду для расчета прогноза в отношении 10 000 преступлений.

Описываемая модель построена в феврале 2021 года. В июле 2021 года, в целях проверки точности построенной модели на данных, которые ранее модель «не видела» и которые не входили ни в обучающий, ни в проверочный датасеты, с помощью построенной модели рассчитана раскрываемость для 16 408 преступлений Приморского края, по которым решение впервые принято в 2021 году – с 01.01.2021 по 30.06.2021. В данный массив вошли переходящий остаток 2020 года и преступления 2021 года – то есть преступления, находившиеся в производстве на конец 2020 года, по которым не было принято никакого решения на 31.12.2020, но по которым к 30.06.2021 принято какое-либо решение, а также преступления, зарегистрированные в январе–июне 2021 года, по которым к 30.06.2021 принято какое-либо решение.

В указанной выборке (16 408 преступлений) соотношение долей раскрытых к нераскрытым преступлениям составило 1,04:1 (8369:8039). По результатам прогнозирования и проверки результатов получены следующие метрики точности построенной модели: Accuracy = 88,64 %, F-мера = 89,37 %. На сбалансированной выборке преступлений вышеуказанной категории (16 078 преступлений) с соотношением долей, раскрытых к нераскрытым преступлениям 1:1, метрики точности составили: Accuracy = 88,49 %, F-мера = 89,05 %, что немного меньше (на 2,28 % для Accuracy) метрик точности по *сбалансированной верификационной* выборке (40 562 преступлений), использовавшейся в процессе обучения.

Таким образом, можно констатировать, что в 9 из 10 случаев, на основе сведений, известных лишь на момент регистрации преступления, построенная модель дает верный прогноз – «преступление будет раскрыто» или «преступление не будет раскрыто».

Отдельно, по итогам 6 месяцев 2021 года, проведена ревизия и рассчитана оценка вероятности раскрываемости 16 534 нераскрытых по итогам 2020 года преступлений, совершенных на территории Приморского края – преступлений, пополнивших массив нераскрытых ППЛ. Данные преступления не использовались в процессе обучения модели и не входили ни в обучающий (280 000), ни в проверочный (60 929) датасеты.

Приостановление в 2020 году производства по данным 16 534 преступлениям произведено по следующим основаниям: в отношении 16 222 преступлений – по п.1 ч.1 ст.208 УПК РФ (лицо, совершившее преступление, не установлено); в отношении 269 преступлений – по п.2 ч.1 ст.208 УПК РФ (лицо установлено и объявлено в розыск); в отношении 43 преступлений – по п.3 ч.1 ст.208 УПК РФ (лицо установлено, его место нахождения известно, однако реальная возможность его участия в уголовном деле отсутствует).

В отношении каждого из 16 534 преступлений сделан прогноз о перспективах их раскрываемости с помощью построенной модели. Данный прогноз сопоставлен с информацией о фактически принятых решениях по данным преступлениям за период с 01.01.2021 по 30.06.2021, которые содержались в статистических карточках Формы № 1.1.

Получены следующие результаты сопоставления:

- из 333 преступлений (2 % от 16 534) со спрогнозированной вероятностью раскрытия большей или равной 50 % фактически раскрыто 61 преступление (18,32 % от 333);
- из 16 201 преступления (98 % от 16 534) со спрогнозированной вероятностью раскрытия меньшей 50 % фактически раскрыто 346 преступлений (2,14 % от 16 201).

Таким образом, удельный вес действительно раскрытых преступлений из числа «перспективных» ППЛ (с вероятностью раскрытия большей или равной 50%) в 8,56 раз выше (18,32 % / 2,14 %), чем у ППЛ, классифицированных моделью как «неперспективные» (с вероятностью раскрытия менее 50 %).

Отдельно следует рассмотреть 269 из 16 534 ППЛ, приостановленные в 2020 году по п.2 ч.1 ст.208 УПК РФ, когда «подозреваемый или обвиняемый скрылся от следствия либо место его нахождения не установлено по иным причинам».

Фактически, лицо, совершившее данное преступление, уже установлено. Необходимо его разыскать и привлечь к уголовной ответственности, чтобы преступление было раскрыто в процессуальном смысле. Казалось бы, модели будет «тяжело понять» на основании лишь сведений из статистической карточки Формы № 1, какие из преступлений, приостановленных «за розыском», имеют благоприятную перспективу на раскрытие.

Однако, если соотнести прогноз, который «сделала» модель по таким преступлениям, с фактически принятыми решениями, то можно увидеть следующее.

Из 269 ППЛ 86 вошли в число 333 ППЛ с «благоприятным» прогнозом на раскрытие, оставшиеся 183 ППЛ вошли в 16 201 ППЛ с «неблагоприятным» прогнозом на раскрытие.

Из 86 преступлений с «благоприятным» прогнозом 52 преступления (60,5 % от 86) за период с 01.01.2021 по 30.06.2021 раскрыты (43) или еще не раскрыты, но по ним возобновлено производство (9).

Из 183 преступлений с «неблагоприятным» прогнозом 64 преступления (35,0 % от 183) за период с 01.01.2021 по 30.06.2021 раскрыты (57) или еще не раскрыты, но по ним возобновлено производство (7).

Таким образом, можно констатировать, что фактически, вероятность раскрытия преступлений, приостановленных «за розыском» из числа преступлений с «благоприятным» прогнозом, в 1,73 раза выше (60,5 % / 35,0 %), чем для преступлений с «неблагоприятным» прогнозом. Получается, что даже здесь модель «чувствует», какие из преступлений, приостановленных за розыском, имеют лучшую перспективу на раскрытие.

Отдельно рассмотрим статьи УК РФ, предусматривающие уголовную ответственность за повторное совершение административного правонарушения лицом, подвергнутым административному наказанию или неоднократное совершение правонарушения изначально предполагают наличие установленного лица, совершившего преступление – это статьи 116.1, 158.1, 264.1 УК РФ, 151.1, 154, 157, 171.4, 180, 212.1, 215.4, 314.1. Преступления, квалифицируемые по данным статьям, не приостанавливались по п.1 ч.1 ст.208 УПК РФ, когда «лицо, подлежащее привлечению в качестве обвиняемого, не установлено».

В число 16 534 ППЛ, квалифицируемым по данным статьям УК РФ, вошло 51 преступление, приостановленное по п.2 или п.3 ч.1 ст.208 УПК РФ, в том числе: ст.157 – 3, ст.264.1 – 36, ст.314.1 – 12. Для 51 преступления данной категории благоприятный прогноз (вероятность более 50%) на раскрытие был сделан моделью в отношении 29

преступлений, неблагоприятный прогноз (вероятность менее 50%) – в отношении 22 преступлений.

Из 29 преступлений с благоприятным прогнозом фактически раскрыто (14) или возобновлено производство (3) по 17 преступлениям; из 22 преступлений с неблагоприятным прогнозом фактически раскрыто (14) или возобновлено производство (0) по 14 преступлениям. Здесь мы получили сопоставимые результаты раскрытия, которые даже несколько больше по преступлениям с неблагоприятным прогнозом – раскрыто или возобновлено производство в отношении 58,6 % (17/29) преступлений с благоприятным прогнозом и раскрыто или возобновлено производство в отношении 63,6 % преступлений с неблагоприятным прогнозом (14/22).

Таблица 1 – Наиболее значимые признаки и степень их влияния на прогноз
Table 1 – The most feature importance

| № п/п | Признак | Значимость, % |
|---|---|---|
| 1 | Статья, часть статьи УК РФ (реквизит 13) | 14,53 |
| 2 | Источник информации о преступлении (реквизит 9.1) | 8,55 |
| 3 | Год возбуждения уголовного дела (реквизит 11) | 7,98 |
| 4 | Дополнительная характеристика № 1 преступления по справочнику № 15 (реквизит 27) | 4,66 |
| 5 | В чьём производстве находится уголовное дело, орган (реквизит 40) | 4,48 |
| 6 | Кем выявлено преступление (реквизит 9) | 4,21 |
| 7 | Место совершения преступления по ОКАТО (реквизит 19.1) | 4,17 |
| 8 | Орган регистрации (расследования) преступления (реквизит 3) | 4,07 |
| 9 | Место совершения преступления по справочнику № 2 (реквизит 19) | 3,94 |
| 10 | Дополнительная характеристика № 2 преступления по справочнику № 15 (реквизит 27) | 3,64 |
| 11 | Номер уголовного дела (реквизит 3) | 3,34 |
| 12 | Тяжесть преступления (реквизит 15) | 2,72 |
| 13 | Признак приготовления или покушения на преступление по ст. 30 УК РФ (реквизит 16) | 2,12 |
| 14 | Размер совершения преступления или причиненного ущерба (реквизит 17) | 1,89 |
| 15 | Результаты осмотра места происшествия (реквизит 29.1) | 1,55 |
| 16 | Уголовное дело возбуждено при отмене постановления об отказе в возбуждении уголовного дела или после отмены постановления о возбуждении уголовного дела (реквизит 10.1) | 1,40 |
| 17 | Характеристика потерпевших (реквизит 33) | 1,38 |
| | **ИТОГО:** | **74,63** |

В тоже время, необходимо учитывать, что построенная модель явным образом не обучалась на прогноз раскрываемости именно ППЛ, когда производство по ранее приостановленному преступлению может быть возобновлено и преступление может быть раскрыто. В этой части видится целесообразным дальнейшее совершенствование модели.

Фреймворк CatBoost позволяет оценить значимость каждого из признаков модели и степень их влияния на прогноз. Весь прогноз на 100% зависит от 83 признаков. В Таблице 1 приведены 17 (20,5 % от 83) наиболее значимых признаков – реквизитов или

составной части реквизитов Формы № 1 – *верификационного* датасета от совокупности которых прогноз зависит на 74,63 %.

## Заключение

Применив современные методы машинного обучения, реализованные в фреймворке CatBoost, удалось создать и обучить прогностическую модель раскрываемости преступлений, совершенных на территории Приморского края. Построенная модель позволяет с высокой степенью точности – более 88 %, используя формализованные и структурированные сведения из статистической карточки Формы № 1, заполняемой на момент регистрации преступления, строить прогноз о том, будет преступление раскрыто или не будет.

Модель позволяет проводить ревизию нераскрытых преступлений прошлых лет в целях определения преступлений, имеющих благоприятную перспективу на раскрытие.

По итогам создания модели удалось определить признаки, в наибольшей степени влияющие на прогноз раскрываемости преступлений.

Планируется продолжить совершенствование модели путем создания дополнительных «синтетических» (расчетных) признаков на базе имеющихся признаков, которые могут повысить точность прогноза и перейти от прогностической модели одного региона к созданию федеральной модели, обученной на федеральном датасете, и позволяющей строить прогноз раскрываемости для преступлений, совершенных на всей территории Российской Федерации.

## ЛИТЕРАТУРА

1. Форма федерального статистического наблюдения № 4-ЕГС «О состоянии преступности и результатах расследования преступлений». Доступно по: http://crimestat.ru/analytics (дата доступа 20.05.2021).
2. Низамов В.Ю. К вопросу о понятии «раскрытие преступления» в криминалистике и уголовном процессе. *Ленинградский юридический журнал*. 2016;1(43):170-179.
3. Приказ Генеральной прокуратуры Российской Федерации, МВД России, МЧС России, Минюста России, ФСБ России, Минэкономразвития России, ФСКН России от 29.12.2005 № 39/1070/1021/253/780/353/399 «О едином учете преступлений». Доступно по: https://rg.ru/2006/01/25/uchet-prestupleniy-dok.html (дата обращения: 20.05.2021).
4. Овчинский В.С., Ларина Е.С. *Искусственный интеллект: Большие данные. Преступность*. Москва: Книжный мир, 2018:1-416.
5. Ясницкий Л.Н. и др. Использование методов искусственного интеллекта в изучении личности серийных убийц. *Криминологический журнал Байкальского государственного университета экономики и права*. 2015;9(3):423-430.
6. Пьянков Д.Д., Малюгин М.И., Ясницкий Л.Н. Применение нейросетевых технологий в изучении факторов, влияющих на преступность в городах России // *Искусственный интеллект в решении актуальных социальных и экономических проблем XXI века*, 21-23 мая 2019 года. Пермь: Пермский государственный национальный исследовательский университет, 2019:126-132.
7. Попонина А.О. Прогнозирование уровня преступности в регионах России // *Искусственный интеллект в решении актуальных социальных и экономических проблем XXI века*, 21-23 мая 2019 года. Пермь: Пермский государственный национальный исследовательский университет, 2019:119-125.
8. Булгаков Д.Ю. Современные подходы к тестированию систем биометрической идентификации по изображению лица. *Искусственный интеллект (большие данные)*

*на службе полиции. Сборник статей международной научно-практической конференции.* Москва: Академия управления МВД России, 2020:45-51.
9. Гордеев А.Ю. Перспективы развития и использования искусственного интеллекта и нейросетей для противодействия преступности в России (на основе зарубежного опыта). *Научный портал МВД России*. 2021;1(53):123-135.
10. Prokhorenkova L., Gusev G., Vorobev A., Dorogush A. V., Gulin A. Catboost unbiased boosting with categorical features. 2017. arXiv: 1706.09516.
11. Melnikov A.V., Narushev I.R., Kubasov I.A. Method for Evaluating Inhomogeneous Alternatives with the Hierarchical Structure of Unrelated Criteria Based on Medium-Consistent Matrix of Pair Comparisons. *Journal of Computational and Engineering Mathematics*. 2019;6(2):32-41. DOI: 10.14529/jcem190203.
12. Салахутдинова К.И., Лебедев И.С., Кривцова И.Е. Алгоритм градиентного бустинга деревьев решений в задаче идентификации программного обеспечения. *Научно-технический вестник информационных технологий, механики и оптики.* 2018:1016-1022. DOI: 10.17586/2226-1494-2018-18-6-1016-1022.
13. Соколов Е. Семинары по выбору моделей. 2015:1-9. Доступно по: http://www.machinelearning.ru/wiki/images/1/1c/Sem06_metrics.pdf (дата обращения: 20.05.2021).

## REFERENCES

1. The form of the federal statistical monitoring No. 4-EGS “On the state of crime and the results of the investigation of crimes”. Available at: http://crimestat.ru/analytics (accessed 20.05.2021). (In Russ)
2. Nizamov V.Y. On the question of the concept of “crime disclosure” in criminalistics and criminal procedure. *Leningrad legal journal*. 2016;1(43):170-179. (In Russ)
3. Order of the General Prosecutor's Office of the Russian Federation, the Ministry of Internal Affairs of Russia, the Ministry of Emergency Situations of Russia, the Ministry of Justice of Russia, the Federal Security Service of Russia, the Ministry of Economic Development of Russia, the Federal Drug Control Service of Russia dated December 29, 2005 No. 39/1070/1021/253/780/353/399 “About the unified accounting of crimes”. Available at: https://rg.ru/2006/01/25/uchet-prestupleniy-dok.html (accessed: 20.05.2021). (In Russ)
4. Ovchinsky V.S., Larina E.S. *Artificial intelligence: Big Data. Crime*. – Moscow: Book World, 2018:1-416. (In Russ)
5. Yasnitsky L.N. et al. The use of artificial intelligence methods in the study of the personality of serial killers. *Criminological journal of the Baikal State University of Economics and Law*. 2015;9(3):423-430. (In Russ)
6. Pyankov D.D., Malyugin M.I., Yasnitsky L.N. The use of neural network technologies in the study of factors affecting crime in Russian cities. *Artificial intelligence in solving urgent social and economic problems of the XXI century*, May 21-23, 2019. Perm: Perm State National Research University. 2019:126-132. (In Russ)
7. Poponina A.O. Forecasting the crime rate in the regions of Russia. *Artificial intelligence in solving urgent social and economic problems of the XXI century,* May 21-23, 2019. Perm: Perm State National Research University, 2019:119-125. (In Russ)
8. Bulgakov D.Y. Modern Approaches to Testing Biometric Identification Systems Based on Facial Images. *Artificial Intelligence (Big Data) in The Service of The Police.* Moscow: Management Academy of the Ministry of the Interior of Russia, 2020:45-51. (In Russ)

9. Gordeev A.Y. Prospects for the development and use of artificial intelligence and neural networks to counter crime in Russia (based on foreign experience). *Scientific portal of the Russian Ministry of internal Affairs.* 2021;1(53):123-135. (In Russ)
10. Prokhorenkova L., Gusev G., Vorobev A., Dorogush A. V., Gulin A. Catboost unbiased boosting with categorical features. 2017. arXiv: 1706.09516.
11. Melnikov A.V., Narushev I.R., Kubasov I.A. Method for Evaluating Inhomogeneous Alternatives with the Hierarchical Structure of Unrelated Criteria Based on Medium-Consistent Matrix of Pair Comparisons. *Journal of Computational and Engineering Mathematics*. 2019;6(2):32-41. DOI: 10.14529/jcem190203.
12. Salakhutdinova K.I., Lebedev I.S., Krivtsova I.E. Algorithm of gradient boosting of decision trees in the problem of software identification. *Scientific and technical bulletin of information technologies, mechanics and optics*. 2018:1016-1022. DOI: 10.17586/2226-1494-2018-18-6-1016-1022. (In Russ)
13. Sokolov E. Seminars on the choice of models. 2015:1-9. Available at: http://www.machinelearning.ru/wiki/images/1/1c/Sem06_metrics.pdf (accessed: 20.05.2021). (In Russ)

## ИНФОРМАЦИЯ ОБ АВТОРАХ / INFORMATION ABOUT AUTHORS

**Булгаков Дмитрий Юрьевич,** первый заместитель начальника Федерального казенного учреждения «Главный информационно-аналитический центр Министерства внутренних дел Российской Федерации», Москва, Российская Федерация.
*e-mail: dbulgakov7@yandex.ru*
ORCID: 0000-0002-9691-6587

**Dmitry Yurevich Bulgakov,** First deputy Chief of Federal government institution «The Main Informational Analytic Centre of the Ministry of Internal Affairs of the Russian Federation», Moscow, Russian Federation.