3

УДК 004.891.3: 005.332.7 **DOI** 10.24412/2413-7383-2025-2-37-66-78

А. В. Звягинцева, И. Ю. Ковалев
Федеральное государственное бюджетное образовательное учреждение
высшего образования «Донецкий государственный университет»
283001, г. Донецк, ул. Университетская, 24

АНАЛИЗ МОДЕЛЕЙ КЛАССИФИКАЦИИ ДЛЯ РАСПОЗНАВАНИЯ ПРЕЦЕДЕНТНЫХ СОБЫТИЙ В ТЕХНОЛОГИЧЕСКИХ ПРОЦЕССАХ ДОБЫЧИ НЕФТИ И ГАЗА

A. V. Zviagintseva, I. Y. Kovalev
Federal State Educational Institution of Higher Education «Donetsk State University»
283001, Donetsk, University str, 24

ANALYSIS OF CLASSIFICATION MODELS FOR RECOGNIZING PRECEDENT EVENTS IN OIL AND GAS PRODUCTION PROCESSES

Г. В. Звягінцева, І. Ю. Ковальов
Федеральна державна бюджетна освітня установа
вищої освіти «Донецький державний університет»
283001, м. Донецьк, вул. Університетська, 24

АНАЛІЗ МОДЕЛЕЙ КЛАСИФІКАЦІЇ ДЛЯ РОЗПІЗНАВАННЯ ПРЕЦЕДЕНТНИХ ПОДІЙ У ТЕХНОЛОГІЧНИХ ПРОЦЕСАХ ВИДОБУТКУ НАФТИ І ГАЗУ

В статье проводится сравнительное тестирование моделей классификации событий на примере датасета аварий на нефтяных скважинах с естественным притоком нефти и газа. Осуществлен анализ и обработка данных, проведена настройка гиперпараметров каждой модели, выполнено обучение моделей, проведено тестирование и рассчитаны метрики.
Ключевые слова: события, нефтедобыча, бинарная классификация, машинное обучение.

The article provides a comparative testing of event classification models using a dataset of accidents at oil wells with natural oil and gas inflow as an example. The data were analyzed and processed, the hyperparameters of each model were adjusted, the models were trained, tested, and metrics were calculated.
Keywords: events, oil production, binary classification, machine learning.

У статті проводиться порівняльне тестування моделей класифікації подій на прикладі датасету аварій на нафтових свердловинах із природним припливом нафти та газу. Здійснено аналіз та обробку даних, проведено налаштування гіперпараметрів кожної моделі, виконано навчання моделей, проведено тестування та розраховано метрики.
Ключові слова: події, нафтовидобуток, бінарна класифікація, машинне навчання.

Нефть играет важнейшую роль в мировой экономике, оказывая влияние на транспорт, производство электроэнергии, нефтехимическую промышленность и национальную безопасность. Главными странами по экспорту нефти в 2023 году являлись Саудовская Аравия (349,1 млн тонн), Россия (240,8 млн тонн), Канада (207,2 млн тонн), США (185 млн тонн) и Ирак (184,2 млн тонн) [1]. В соответствии с Энергетической стратегией России [2] нефтегазовая отрасль должна обеспечить стабильную добычу нефти с газовым конденсатом в объеме 525 млн т в год, увеличение коэффициента извлечения нефти до 40%, переход на новую технологическую платформу в освоении трудноизвлекаемых запасов, малых месторождений, малодебитных и высокообводненных скважин.

Нефтегазоразведка и добыча сопряжены с многочисленными проблемами, среди которых выделяются риски промышленной и экологической безопасности [3].

Аварии, такие как взрывы на платформах или разливы нефти, могут привести к серьезным экологическим катастрофам и человеческим жертвам. Своевременное распознавание опасных событий на основе данных технологического мониторинга, а также выявление прецедентов по ретроспективным данным – актуальная задача для многих отраслей промышленности.

Цель работы – провести сравнительный анализ моделей классификации для выявления прецедентов на примере датасета аварий на нефтяных скважинах с естественным притоком нефти и газа.

Опасные события в нефтегазовой отрасли

В Российской Федерации за период 2004–2018 гг. произошла 251 аварийная ситуация на объектах нефтепереработки и нефтехимии. Наиболее частыми причинами произошедших аварий являлись взрыв (41%), выброс опасных веществ (20%) и пожар (39%). Общий экономический ущерб за рассмотренный период составил 20 млрд 863 млн руб. [4]. В свою очередь, за период 2019–2023 гг. на объектах нефтеперерабатывающих предприятий в России произошло 53 аварий, на которых погибло 30 человек. Основная доля аварий пришлась на пожар (58%) и взрыв (26%), оставшиеся 16% – выброс опасных веществ. Также за последние 10 лет в РФ произошло как минимум 5 аварий на нефтескважинах, которые удалось бы предотвратить, применяя системы мониторинга и системы раннего обнаружения неисправностей [5].

Раннее обнаружение и диагностика неисправностей при технологическом мониторинге помогает избежать аномального развития событий и снизить репутационные, экономические и экологические потери. Поэтому решению этой проблемы уделяется большое внимание.

В данной работе исследуются технологические аварии на примере нефтедобывающих скважин с естественным притоком нефти и газа. Такие аварии характеризуются множеством разнородных факторов и сложными причинно-следственными связями инициирующих событий.

Классификация аварий на нефтяных скважинах регламентируется национальными и международными стандартами: ГОСТ Р 53554-2009, ГОСТ Р 22.0.05-94, РД 08-492-02, API RP 59, API STD 53, ISO 10418:2003, ISO 14224:2016, IOGP Report 476, а также методическими рекомендациями [6]. Типы аварий на нефтяных скважинах делятся по характеру возникновения: внезапные (выбросы, фонтанирование), постепенные (утечки, коррозия); по масштабу последствий: локальные (в пределах скважины), объектовые (на территории месторождения), региональные (загрязнение окружающей среды), глобальные (катастрофические разливы); по виду осложнений:

газонефтеводопроявления (ГНВП), открытые фонтаны, пожары и взрывы, обрушения ствола скважины, разрушение устья; по причинам: технические (поломка оборудования), технологические (ошибки в бурении/эксплуатации), природные (землетрясения, паводки). Исходя из технологических особенностей, признаков реализации опасности аварий, тяжести последствий по трем уровням опасности возникновения, развития и эскалации аварий различают: 1-й уровень: чрезвычайно высокая аварийная опасность – авария; 2-й уровень: высокая аварийная опасность – инцидент; 3-й уровень: средняя аварийная опасность – предпосылка к инциденту [6].

В качестве события обычно рассматривается любой наблюдаемый факт, который выражается в изменении состояния объекта или системы; факт, который привлекает внимание нарушением технологических режимов. В свою очередь, в качестве прецедента изучается событие, которое имело место в прошлом и является примером и основанием для аналогичных действий в настоящем. События, являющиеся прецедентными и повлёкшие за собой ряд других событий, оказывают значительное влияние на формирование опасных ситуаций. В контексте нефтегазовой добычи прецедентные события играют критическую роль. Например, к ним относятся аварии на буровых платформах, утечки нефти и газа, взрывы на газопроводах, а также сбои в работе оборудования, приводящие к остановке добычи нефти и газа. Анализ этих событий позволяет выявлять недостатки в системе, разрабатывать меры по предотвращению повторения подобных ситуаций и повышать безопасность и эффективность добычи [7].

Согласно Цюрихской схеме классификации [8] к технологическим и техническим событиям (технологическим рискам) относятся технологические ошибки и проблемы, сбои технического, программного и аппаратного обеспечения, инциденты в области промышленной, информационной и технической безопасности, аварии и т.д. В модели управлении рисками организаций COSO [9] технологические события – это события, связанные с технологическими новшествами, техническим прогрессом, использованием инфраструктуры, спросом на продукты и услуги, ведущие к снижению затрат и повышению эффективности деятельности и т.д. Методы выявления и оценки событий в модели COSO являются преимущественно экспертными, используется методология деревьев событий/решений, риски и возможности просчитываются с учетом взаимосвязи событий и их вероятностей [7], [10].

Прецедентный подход [11–13] позволяет упростить процесс принятия решений в условиях временных ограничений и при наличии различного рода неопределенности в исходных данных и экспертных знаниях [14], [15]. Анализ прецедентных событий применяется для раннего определения аварийных (нештатных, непредвиденных, случайных) событий с целью предотвращения аварий и требует использования ретроспективных данных [16].

Технология подводной добычи нефти и газа

В работе в качестве примера рассматриваются морские скважины так как они являются одним из наиболее сложных и технологически емких типов скважин, требующих высокого уровня технической и экологической безопасности. Морские скважины имеют ряд особенностей, включая высокую стоимость бурения и эксплуатации, сложность доступа к месторождениям и необходимость использования специализированных платформ [17], [18].

В процессе эксплуатации нефтяных скважин накапливается большой объем данных мониторинга, характеризующих режим работы и свойства добываемого сырья, т.к. скважина представляет собой набор датчиков, механических, пневматических и

гидравлических систем, которые могут быть частично или полностью установлены на морском дне, в скважине или на поверхности. На рисунке 1 представлена схема морской скважины с естественным притоком нефти и газа [19].

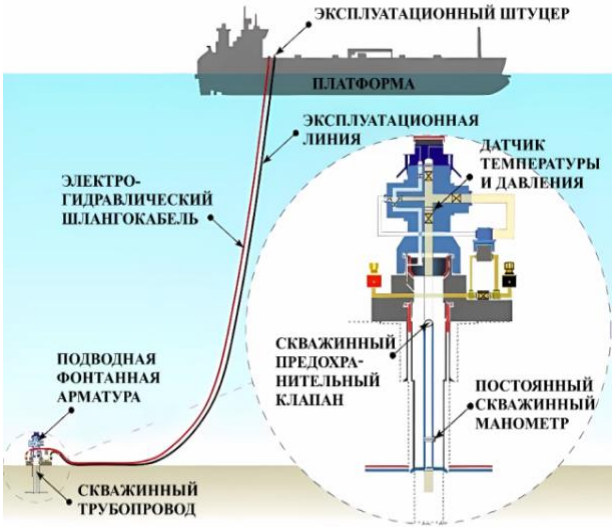

Рисунок 1 – Система подводной добычи нефти и газа: *источник:* [20, с. 3]

Как видно из рисунка, нефть или газ поступают из трубопровода скважины через подводную фонтанную арматуру (ПФА) и гидравлический шланг в емкости платформы под давлением, не более 103,5 МПа при температуре не выше 121 °С. Управление потоком добываемой среды обеспечивается автоматизированной ПФА, которая размещается на морском дне у устья скважины. Управление регулирующим клапаном и датчиками производится с поверхности через шлангокабель, представляющий собой комплекс электрических и волоконно-оптических кабелей, шлангов или трубопроводов, заключенных в общую оболочку.

Шлангокабель предназначен для передачи сигналов связи, электрической и гидравлической энергии, а также химических сред [21]. Внутри ПФА установлен постоянный скважинный манометр, скважинный предохранительный клапан (СПК), датчик температуры и давления. В целом описанная система представляет собой автоматизированный технологический комплекс, обеспечивающий управление по всей технологической цепочке добычи нефти и газа [21].

Опасные события и ситуации, такие как аварии или отказы оборудования, могут возникнуть в процессе добычи и требуют немедленного реагирования. Обычно на возобновление процесса требуется порядка нескольких часов или дней, что в некоторых случаях может привести к длительным простоям и значительным экономическим потерям.

Исходные данные и их первичная обработка

В качестве датасета использован набор данных из открытого репозитория на GitHub [22], представленный девятью классами событий: 0 класс – нормальная работа; 1–8 классы – опасные события и технологические сбои.

1 класс. Резкое увеличение содержания взвешенной воды, отложений и других примесей в продукции. Показатель определяется как отношение между скоростью потока воды (осадков) и скоростью потока жидкости, измеренных при нормальной температуре и давлении. В течение жизненного цикла скважины показатель увеличивается за счет естественного водоносного горизонта пласта, либо из-за искусственной закачки. Внезапное увеличение показателя приводит к проблемам с переработкой нефти и к снижению коэффициента нефтеотдачи.

2 класс. Ложное срабатывание/закрытие СПК часто происходит без каких-либо признаков на поверхности (например, падение давления в гидравлическом приводе). Приводит к незапланированной остановке добычи.

3 класс. Воздушная пробка. Критический тип нестабильности. Две наиболее яркие особенности этого события – четко выраженная периодичность (около 30, 45 или 60 мин. неблагоприятного события) и интенсивность, которая обычно достаточна для обнаружения датчиками по всей производственной линии. События такого типа могут привести к повреждению оборудования скважины и/или всего добывающего комплекса.

4 класс. Нестабильность потока. Во время нестабильности потока одна из контролируемых переменных претерпевает соответствующие изменения, но с допустимыми амплитудами. Характеристикой, которая отличает этот тип от воздушных пробок, является отсутствие периодичности между этими изменениями. Нестабильность потока может прогрессировать до воздушной пробки.

5 класс. Быстрое снижение производительности. Производительность скважины с естественным притоком зависит от статического давления в резервуаре, процентного содержания основного осадка и воды, вязкости добываемой жидкости, диаметра эксплуатационной линии и т.д. Когда энергии системы становится недостаточно для преодоления потерь, поток среды замедляется или даже прекращается.

6 класс. Резкое ограничение клапаном эксплуатационного штуцера (ЭШ). Для корректного использования выражения «Резкое ограничение клапаном ЭШ» ограничение должно происходить с амплитудой выше установленного эталона (например, 5%) и в течение короткого времени (например, менее 10 с). Краткие сбои могут наблюдается при ручном управлении из-за эксплуатационных проблем.

7 класс. Образование отложений в эксплуатационном штуцере (ЭШ). Мониторинг производственного клапана важен из-за восприимчивости к неорганическим отложениям, которые могут значительно снизить добычу нефти и газа.

8 класс. Гидраты в производственной линии. Кристаллогидраты могут образовываться в скважинах, приводя к полному прекращению потока.

В репозитории на GitHub доступны только небольшие фрагменты, в которых состояние переходит от обычной (нормальной) работы к переходному (предаварийному) состоянию, а затем к установившейся аномалии (аварии). Полученные фрагменты данных в формате CSV помещали в каталог, который соответствует одному из зафиксированных классов событий в морских скважинах с естественным притоком нефти и газа. Количество фрагментов (файлов) для каждого класса события представлено в таблице 1.

Таблица 1 – Количество файлов для каждого класса события датасета аварий на нефтяных скважинах с естественным притоком нефти и газа

Класс события	Количество фрагментов (файлов)			
	Реальные	Симуляция	Экспертные	Всего
0 – нормальная работа	594	0	0	594
1 – резкий рост содержания взвешенной воды, отложений и др. примесей в продукции	5	114	10	129
2 – ложное срабатывание СПК	20	16	0	36
3 – серьёзная закупорка	32	74	0	106
4 – нестабильность потока	344	0	0	344
5 – быстрое снижение производительности	11	439	0	450
6 – резкое ограничение клапаном ЭШ	6	215	0	221
7 – образование отложений в ЭШ	5	0	10	15
8 – гидраты в производственной линии	0	81	0	81
ИТОГО	1017	939	20	1976

Из таблицы 1 видно, что с 2014 по 2018 год по 18-ти нефтяным платформам произошло не так много аварий, поэтому сохранено всего 423 файла с аварийными состояниями. Для дальнейших исследований использовались только реальные данные.

На протяжении всего фрагмента значения показателей (данные датчиков) фиксировались ежесекундно. Объем данных в файлах зависит от продолжительности аварии и находится в диапазоне от 3 673 до 345 601 строк. Минимальное количество данных в файле наблюдается для второго класса событий, а максимальное – для седьмого. Следует отметить, что каждый из файлов имеет одинаковую структуру, её можно представить в виде таблицы 2.

Таблица 2 – Описание показателей датасета аварий на нефтяных скважинах с естественным притоком нефти и газа

№	Показатель	Описание	Ед. изм.
1	Timestamp	Временная метка	с
2	P-PDG	Давление в датчике давления на выходе из скважины	Па
3	P-TPT	Давление в датчике давления и температуры возле устья скважины	Па
4	T-TPT	Температура в датчике давления и температуры возле устья скважины	°C
5	P-MON-CKP	Давление перед затвором регулирующего клапана	Па
6	T-JUS-CKP	Температура за затвором регулирующего клапана	°C
7	P-JUS-CKGL	Давление на клапане газовой линии	Па
8	T-JUS-CKGL	Температура на клапане газовой линии	°C
9	QGL	Скорость потока газа	см³/с
10	Class	Состояние процесса: нормальное (0), предаварийное (101, 102, 103, 104, 105, 106, 107, 108), аварийное (1, 2, 3, 4, 5, 6, 7, 8)	–

Показатель, качественно характеризующий состояние процесса, разбитый на нормальное (0), предаварийное (101, 102, 103, 104, 105, 106, 107, 108) и аварийное (1, 2, 3, 4, 5, 6, 7, 8) состояния, позволяет четко выделить класс события и вовремя принять нужные меры по устранению неисправности.

Для нулевого класса в репозитарии [22] представлено 9 903 750 значений измерений, для 1-ого – 9 182 499, для 2-ого – 655 017, для 3-его – 4 838 079, 4-ого – 2 462 076, 5-ого – 13 414 798, 6-ого – 5 861 368, для 7-ого и 8-ого – 2 885 548 и 2 186 920 соответственно. Таким образом, общий объем исходных данных составил 51 390 055 наблюдений.

Проведен анализ значений количественных переменных (признаков), выявлено, что большинство значений P-PDG, T-JUS-CKGL, QRL равны нулю и не несут полезную информацию, поэтому эти переменные были исключены из датасета. В результате оставлено 5 наиболее информативных признаков: P-TPT, T-TPT, P-MON-CKP, T-JUS-CKP, P-JUS-CKGL.

В данных наблюдался большой дисбаланс по количеству событий для каждого класса, поэтому значения качественной переменной Class были приведены к бинарному виду: 1 – нормальное состояние (74%), 0 – предаварийное и аварийное (26%). Датасет разбили на обучающую (80% все данных) и тестовую (20%) выборки.

Модели классификации

На сегодняшний день для решения задачи классификации разработано большое количество алгоритмов и методов, на основе которых строятся модели классификации (табл. 3). В данной работе рассмотрены все методы из этой таблицы. Они реализованы

в открытой библиотеки Scikit-learn – одной из наиболее широко используемых библиотек Python для Data Science и Machine Learning. Библиотека содержит алгоритмы машинного обучения: классификации, прогнозирования или разбивки данных на группы, была разработана в рамках проекта Google Summer of Code в 2007 году, а сейчас активно развивается и поддерживается.

Таблица 3 – Сравнение методов классификации по разным критериям

Критерий	Группы методов
Тип обучения	**Без учителя (Unsupervised / Anomaly Detection):** One Class SVM, Local Outlier Factor, Isolation Forest **С учителем (Supervised):** Linear SVM, Logistic Regression, RBF SVM, Naive Bayes, Neural Net, QDA, AdaBoost, Nearest Neighbors, CatBoost, Histogram-Based Gradient Boosting, Decision Tree, XGBoost, Extra Trees, Random Forest, Bagging Classifier
Основная задача	**Обнаружение аномалий:** One Class SVM, Local Outlier Factor, Isolation Forest **Классификация:** Linear SVM, Logistic Regression, RBF SVM, Naive Bayes, Neural Net, QDA, AdaBoost, Nearest Neighbors, CatBoost, Histogram-Based GB, Decision Tree, XGBoost, Extra Trees, Random Forest, Bagging Classifier
Алгоритмическая группа	**Методы, основанные на расстоянии / ядрах:** SVM (Linear, RBF, One Class), Nearest Neighbors, Local Outlier Factor **Вероятностные методы:** Naive Bayes, QDA, Logistic Regression **Изоляция / разделение:** Isolation Forest **Методы на основе деревьев:** Decision Tree, Random Forest, Extra Trees, XGBoost, CatBoost, Histogram-Based GB, AdaBoost (может использовать деревья) **Ансамблевые методы:** AdaBoost, Bagging Classifier, Random Forest, Extra Trees, XGBoost, CatBoost, Histogram-Based GB **Нейросетевые методы:** Neural Net
Чувствительность к выбросам	**Чувствительные:** SVM (без регуляризации), Logistic Regression, Neural Net **Устойчивые:** Isolation Forest, Local Outlier Factor, Random Forest, Extra Trees, Bagging
Интерпретируемость	**Высокая:** Decision Tree, Logistic Regression, Naive Bayes, QDA **Средняя:** Random Forest, AdaBoost, Linear SVM **Низкая:** Neural Net, RBF SVM, CatBoost, XGBoost, Histogram-Based GB
Масштабируемость	**Высокая:** Linear SVM, Logistic Regression, Decision Tree, Random Forest, CatBoost, XGBoost, Histogram-Based GB **Средняя:** AdaBoost, Neural Net (зависит от архитектуры) Низкая: RBF SVM (плохо на больших данных), Nearest Neighbors (плохо в высоких размерностях)
Поддержка многоклассовой классификации	**Да:** Logistic Regression, Random Forest, Neural Net, Naive Bayes, Decision Tree, AdaBoost, XGBoost, CatBoost, Extra Trees, Bagging, RBF SVM (с OvO/OvR) **Нет (только бинарная / аномалии):** Linear SVM (без OvO/OvR), One Class SVM, Local Outlier Factor, Isolation Forest

Используемые метрики

Для оценки качества моделей бинарной классификации будем использовать следующие метрики: *Precision* (точность), *Recall* (полнота), *F1-score* (среднее гармоническое между точностью и полнотой) и *ROC AUC* (площадь под кривой ROC) [16]. Метрика *Precision* (точность) показывает долю реальных объектов класса среди отнесённых классификатором к этому классу:

$$\Pr ecision = TP/(TP + FP), \quad (1)$$

где TP (True Positive) – количество объектов, которые модель верно отнесла к классу, FP (False Positive) – количество объектов, ошибочно отнесенные моделью к классу. Значение этой метрики должно стремиться к единице.

Метрика *Recall* (полнота) оценивает долю правильно классифицированных объектов данного класса:

$$\mathrm{Re}call = TP/(TP + FN), \tag{2}$$

где TP (True Positive) – количество правильно классифицированных объектов, FN (False Negative) – количество объектов, которые модель ошибочно не отнесла к классу. Метрика показывает сколько объектов потеряно при классификации. Чем ближе значение к 1 (100%) тем лучше модель справляется с задачей распознавания.

Для общей оценки качества модели используем показатель *F1-score* (*F1-мера*) – среднее гармоническое значение между точностью и полнотой:

$$F1 - score = 2 \cdot \mathrm{Pr}ecision \cdot \mathrm{Re}call/(\mathrm{Pr}ecision + \mathrm{Re}call), \tag{3}$$

где 1 соответствует идеальной классификации.

Часто результат работы визуализируют с помощью кривой ошибок *ROC* (receiver operating characteristic), иллюстрирующей производительность классификационной модели и отражающей графическое представление компромисса между чувствительностью и специфичностью при различных порогах классификации. Ось X данного графика (FPR) – ложноположительная частота (доля ошибочно классифицированных отрицательных результатов относительно всех отрицательных результатов), а ось Y – истинноположительная частота ответов (*Recall*).

$$FPR = FP/(TN + FP), \tag{4}$$

где FP (False Positive) – количество объектов, ошибочно отнесенных моделью к классу, а TN (True Negative) – количество объектов, которые модель верно не отнесла к классу.

В этом случае качество модели оценивается как *AUC* (Area Under the ROC Curve) – мера, которая позволяет суммировать производительность модели одним числом, измеряя площадь под кривой *ROC*. *AUC* колеблется от 0 до 1. Чем ближе значение к 1 (100%) тем лучше модель справляется с задачей распознавания.

Работа над повышением качества модели сводится к получению более высокой оценки *ROC AUC*, *Precision, Recall*, и, следовательно, *F1-score*. Скорость обучения моделей важна для самообновления и самообучения моделей, так как быстрое обучение приводит к более короткому времени отклика всей системы. Для оценки скорости потоковой работы и для работы на слабых устройствах необходимо знать время выполнение модели на тестовых данных.

Сравнительный анализ моделей классификации

Для обработки и анализа данных, обучения и сравнительного анализа качества моделей классификации использовали язык программирования Python (большое количество библиотек) и Jupyter Notebook для визуализации данных. Все операции проводились в облачном сервисе Yandex DataSphere, который предназначен для анализа данных, разработки и эксплуатации моделей машинного обучения в составе платформы Yandex.Cloud. Анализ качества модели включал подбор гиперпараметров, обучение на тренировочном наборе данных, тестирование на тестовом наборе данных (на вход модели подаётся набор показателей, а модель выдаёт класс события), расчёт метрик. После обучения модель сохраняли с целью последующего использования в качестве основы для написания программного продукта.

Изучаемые модели чувствительны к гиперпараметрам, для определения их оптимальных значений использовались инструменты RandomizedSearchCV и GridSearchCV. GridSearchCV осуществляет полный перебор всех комбинаций параметров, этот инструмент использовался при малом количестве этих величин. Если параметров было много, то применялся RandomizedSearchCV, который выполняет случайный поиск по распределениям параметров. Диапазоны параметров выбирались на основе документации, экспериментов и экспертной оценки. Использовались также стандартные рекомендации из библиотеки sklearn. Для уменьшения переобучения под конкретный параметр данные разбивались на несколько частей, а модель обучалась и тестировалась на разных комбинациях этих частей. Полученные результаты сведены в таблицу 4.

Таблица 4 – Результаты тестирования моделей классификации на примере датасета аварий на нефтяных скважинах

Модель	*Precision* (точность)	*Recall* (полнота)	*F1-score* (*F-мера*)	*ROC AUC*	Время обучения, с	Время тестирования, с
One Class SVM - RBF	0,574	0,579	0,577	0,463	250,00	27,81
Local outlier factor	0,614	0,713	0,624	0,503	1,36	0,30
Isolation Forest	0,784	0,757	0,766	0,740	0,30	0,57
Linear SVM	0,786	0,426	0,392	0,598	9,11	0,01
Logistic Regression	0,549	0,683	0,600	0,475	1,13	0,01
RBF SVM	0,696	0,729	0,620	0,505	1656,54	119,50
Naive Bayes	0,652	0,705	0,661	0,539	0,02	0,00
Neural Net	0,834	0,836	0,822	0,734	14,18	0,11
QDA	0,884	0,881	0,873	0,800	0,06	0,01
AdaBoost	0,953	0,953	0,953	0,935	4,43	0,13
Nearest Neighbors	0,981	0,981	0,981	0,977	0,11	1,66
CatBoost	0,982	0,982	0,982	0,979	4,34	0,02
Histogram-Based Gradient Boosting	0,987	0,987	0,987	0,986	0,92	0,07
Decision Tree	0,988	0,988	0,988	0,985	0,64	0,00
XGBoost	0,989	0,989	0,989	0,986	10,53	0,04
Extra Trees	0,989	0,989	0,989	0,987	2,15	0,24
Random Forest	0,989	0,989	0,989	0,987	42,15	0,44
Bagging Classifier	0,990	0,990	0,990	0,988	3,47	0,02

Из таблицы видно, что при работе с датасетом аварий на нефтяных скважинах с естественным притоком нефти и газа у модели Bagging Classifier метрика *F1-score* достигла 99%, что является отличным результатом для выявления аномального состояния работы нефтяной скважины. Скорость обучения большинства моделей на основе ансамблей решающих деревьев (Decision Tree, Extra Trees, Bagging Classifier) достаточная для периодического самообновления и самообучения моделей, и скорость определения класса события на тестовых данных отличная. А вот скорость обучения моделей SVM и Neural Net низкая, а значит они плохо подходят для потоковой работы в реальных условиях.

Выводы

В ходе проведения сравнительного анализа моделей классификации прецедентных событий на примере датасета аварий на нефтяных скважинах выполнен анализ и обработка данных, проведена настройка гиперпараметров каждой модели при помощи

инструментов GridSearchCV и RandomizedSearchCV, выполнено обучение моделей двух категорий: без учителя (One Class SVM, Local outlier factor, Isolation Forest) и с учителем (Linear SVM, RBF SVM, Naive Bayes, Neural Net, QDA, AdaBoost, Nearest Neighbors, CatBoost, Histogram-Based Gradient Boosting, Decision Tree, XGBoost, Extra Trees, Random Forest, Bagging Classifier), проведено тестирование и рассчитаны метрики *Precision* (точность), *Recall* (полнота), *F1* (*F-мера*), *ROC AUC*, время обучения, время тестирования.

Модели классификации без учителя в теории хорошо подходят для выявления аномалий в неразмеченных данных, однако на практике для классификации прецедентных событий эти модели на примере датасета аварий показали плохие результаты (40–58% правильных ответов). В большей степени это связано с тем, что нормальные экземпляры находятся слишком близко к аномалиям.

Модели классификации с учителем на основе ансамблей деревьев решений (CatBoost, Histogram-Based Gradient Boosting, Decision Tree, XGBoost, Extra Trees, Random Forest, Bagging Classifier) справились с задачей классификации отлично, показав и высокие значения метрик (примерно 99% правильных ответов), и высокую скорость обучения (4 с на переобучение всей модели) и работы (0,02 с на прогноз класса состояния).

Отметим, что алгоритм классификации на основе многослойной нейронной сети считается очень перспективным, но для задачи бинарной классификации показал всего 82% правильных ответов. Такие результаты подтверждают тот факт, что для задачи бинарной классификации лучше деревьев решений пока еще нет алгоритмов.

Полученный опыт можно использовать для обнаружения аварий в других сферах деятельности человека. В этой работе не было уделено должное внимание определению конкретных классов событий (предаварийных и аварийных), чтобы можно было рекомендовать действия в зависимости от класса неисправности. Поэтому необходимо исследовать возможности алгоритмов классификации при мультиклассовой классификации, когда наблюдается большой дисбаланс по классам. Также важно отметить, что данные являются темпоральными, и признак времени играет значительную роль при прогнозирования аварийных ситуаций с помощью алгоритмов регрессии.

Список литературы

1. Energy Institute Statistical Review of World Energy 2024 – Energy Institute. 2024, 76 p.
2. Энергетическая стратегия Российской Федерации до 2050 года. М.: Министерство энергетики РФ, 2025. 107 с.
3. Шмаль Г.И. Проблемы при разработке трудноизвлекаемых запасов нефти в России и пути их решения // Георесурсы, №18(4), 2016. – С. 256–260.
4. Калараш Р.А., Короткова Т.Г. Статистика аварий на объектах нефтехимической и нефтеперерабатывающей промышленности // Научные труды КубГТУ, №7, 2019. – С. 314–324.
5. Уроки, извлеченные из аварий. – Текст: электронный // Ростехнадзор: [сайт]. URL https://www.gosnadzor.ru/industrial/oil/lessons/ (28.05.2025).
6. Руководство по безопасности «Методические рекомендации по классификации аварийно опасных происшествий на опасных производственных объектах нефтегазового комплекса». Утв. Приказом Ростехнадзора от 20.11.2023, №410. – 17 с.
7. Звягинцева А.В., Гучмазова Т.К., Клеменюк В.Р. Выявление взаимосвязи сложных событий на примере анализа статистических данных о чрезвычайных ситуациях // Вестник ДонНУ. Серия Г: Технические науки, №3, 2024. – С. 45–54.
8. Alvarez G. Operational Risk Quantification: Mathematical Solutions for Analyzing Loss Data, 2001, 18 p.
9. Управление рисками организаций. Интегрированная модель. Краткое изложение. Концептуальные основы. Проектный консультативный совет COSO, 2004. – 111 с.
10. Звягинцева А.В. Вероятностные методы комплексной оценки природно-антропогенных систем. – М.: Спектр, 2016. – 258 с.

11. Черновалова М.В., Черненский Л.Л., Макарова М.М. Прецедентный подход для оценки влияния молний на системы уличного освещения с использованием онтологий // Программные продукты и системы, №35(4), 2022. – С. 729–736.
12. Микрюков А.А., Куулар А.В. Совершенствование процесса управления инцидентами на основе прецедентного подхода // Открытое образование. Т.25, №4, 2021. – С. 47–54.
13. Рычка О.В. Анализ эффективности усовершенствованных методов поиска и обработки аномалий для нелинейных моделей с внутренней линейностью // Проблемы искусственного интеллекта, №3(18), 2020. – С. 101–110.
14. Кривов М.В., Асламова Е.А., Асламова В.С. Система выработки стратегий управления промышленной безопасностью // Вестник Томского государственного университета. Управление, вычислительная техника и информатика, №59, 2022. – С. 55–65.
15. Сафонов В.С., Одишария Г.Э., Швыряев А.А. Теория и практика анализа риска в газовой промышленности. – М.: НУМЦ Минприроды России, 1996. – 207 с.
16. David M.W. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation // International Journal of Machine Learning Technology. Vol.2, no1, 2011: 37–63.
17. NORSOK D-010 Well integrity in drilling and well operations. 2013. Standards Norway, 224 p.
18. ГОСТ Р 54483-2021. Нефтяная и газовая промышленность. Сооружения нефтегазопромысловые морские. Общие требования. – М.: Российский институт стандартизации, 2021. – 45 с.
19. ГОСТ Р 53554-2009. Поиск, разведка и разработка месторождений углеводородного сырья. Термины и определения. – М.: Стандартинформ, 2020. – 19 с.
20. R.E.V. Vargas, et al. A realistic and public dataset with rare undesirable real events in oil wells // Journal of Petroleum Science and Engineering. Vol.181, 2019, 9 p.
21. ГОСТ Р 59304-2021. Нефтяная и газовая промышленность. Системы подводной добычи. Термины и определения. – М.: Стандартинформ, 2021. – 16 с.
22. 3W Dataset [online]. Available at: https://github.com/ricardovvargas/3w_dataset (accessed May 20, 2025).

References

1. Energy Institute Statistical Review of World Energy 2024 – Energy Institute. 2024, 76 p.
2. Jenergeticheskaja strategija Rossijskoj Federacii do 2050 goda. Moscow, Ministerstvo jenergetiki Rossijskoj Federacii, 2025, 107 p.
3. Shmal' G.I. Problemy pri razrabotke trudnoizvlekaemyh zapasov nefti v Rossii i puti ih reshenija. *Georesursy*, no18(4), 2016: 256–260.
4. Kalarash R.A., Korotkova T.G. Statistika avarij na ob"ektah neftehimicheskoj i neftepererabatyvajushhej promyshlennosti. *Nauchnye trudy KubGTU*, no 7, 2019: 314–324.
5. Uroki, izvlechennye iz avarij. Tekst: jelektronnyj. Rostehnadzor: [sajt]. URL https://www.gosnadzor.ru/industrial/oil/lessons/ (28.05.2025).
6. Rukovodstvo po bezopasnosti “Metodicheskie rekomendacii po klassifikacii avarijno opasnyh proisshestvij na opasnyh proizvodstvennyh ob"ektah neftegazovogo kompleksa”. Utv. Prikazom Rostehnadzora ot 20.11.2023, no 410, 17 p.
7. Zviaginceva A.V., Guchmazova T.K., Klemenjuk V.R. Vyjavlenie vzaimosvjazi slozhnyh sobytij na primere analiza statisticheskih dannyh o chrezvychajnyh situacijah. *Vestnik DonNU. Serija G: Tehnicheskie nauki*, no 3, 2024: 45–54.
8. Alvarez G. Operational Risk Quantification: Mathematical Solutions for Analyzing Loss Data, 2001, 18 p.
9. Upravlenie riskami organizacij. Integrirovannaja model'. Kratkoe izlozhenie. Konceptual'nye osnovy. Proektnyj konsul'tativnyj sovet COSO, 2004, 111 p.
10. Zviaginceva A.V. Verojatnostnye metody kompleksnoj ocenki prirodno-antropogennyh sistem. Moscow, Spektr, 2016, 258 p.
11. Chernovalova M.V., Chernenskij L.L., Makarova M.M. Precedentnyj podhod dlja ocenki vlijanija molnij na sistemy ulichnogo osveshhenija s ispol'zovaniem ontologij. *Programmnye produkty i sistemy*, no 35(4), 2022: 729–736.
12. Mikrjukov A.A., Kuular A.V. Sovershenstvovanie processa upravlenija incidentami na osnove precedentnogo podhoda. *Otkrytoe obrazovanie*. V.25, no 4, 2021: 47–54.
13. Rychka O.V. Analiz jeffektivnosti usovershenstvovannyh metodov poiska i obrabotki anomalij dlja nelinejnyh modelej s vnutrennej linejnost'ju. *Problemy iskusstvennogo intellekta*, no 3(18), 2020: 101–110.
14. Krivov M.V., Aslamova E.A., Aslamova V.S. Sistema vyrabotki strategij upravlenija promyshlennoj bezopasnost'ju. *Vestnik Tomskogo gosudarstvennogo universiteta. Upravlenie, vychislitel'naja tehnika i informatika*, no 59, 2022: 55–65.
15. Safonov V.S., Odisharija G.Je., Shvyrjaev A.A. Teorija i praktika analiza riska v gazovoj promyshlennosti. Moscow, NUMC Minprirody Rossii, 1996, 207 p.

16. David M.W. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation // International Journal of Machine Learning Technology. Vol.2, no1, 2011: 37–63.
17. NORSOK D-010 Well integrity in drilling and well operations. 2013. Standards Norway, 224 p.
18. GOST R 54483-2021. Neftjanaja i gazovaja promyshlennost'. Sooruzhenija neftegazopromyslovye morskie. Obshhie trebovanija. Moscow, Rossijskij institut standartizacii, 2021, 45 p.
19. GOST R 53554-2009. Poisk, razvedka i razrabotka mestorozhdenij uglevodorodnogo syr'ja. Terminy i opredelenija. Moscow, Standartinform, 2020, 19 p.
20. R.E.V. Vargas, et al. A realistic and public dataset with rare undesirable real events in oil wells // Journal of Petroleum Science and Engineering. . Vol.181, 2019, 9 p.
21. GOST R 59304-2021. Neftjanaja i gazovaja promyshlennost'. Sistemy podvodnoj dobychi. Terminy i opredelenija. Moscow, Standartinform, 2021, 16 p.
22. 3W Dataset [online]. Available at: https://github.com/ricardovvargas/3w_dataset (accessed May 20, 2025).

RESUME

A. V. Zviagintseva, I. Y. Kovalev

Analysis of classification models for recognizing precedent events in oil and gas production processes

Background: The oil and gas industry plays a vital role in the global economy, impacting transportation, power generation, petrochemical industry, and national security. Accidents such as platform explosions or oil spills cause serious environmental disasters and human casualties. Timely recognition of hazardous events based on process monitoring data, as well as identifying precedents based on historical data, is a pressing issue for many industries. The purpose of this paper is to conduct a comparative analysis of classification models for identifying precedents using a dataset of accidents at oil wells with natural oil and gas inflows as an example.

Materials and methods: data analysis and processing were performed, hyperparameters of each model were tuned using the GridSearchCV and RandomizedSearchCV tools, models of two categories were trained: unsupervised (One Class SVM, Local outlier factor, Isolation Forest), supervised (Linear SVM, RBF SVM, Naive Bayes, Neural Net, QDA, AdaBoost, Nearest Neighbors, CatBoost, Histogram-Based Gradient Boosting, Decision Tree, XGBoost, Extra Trees, Random Forest, Bagging Classifier), testing was performed and the following metrics were calculated: *Precision*, *Recall*, *F1 (F-measure*), *ROC AUC*, training time, testing time.

Results: unsupervised classification models are theoretically well suited for detecting anomalies in unlabeled data, but in practice, these models showed poor results (40–58% correct answers) for classifying precedent events using the accident dataset as an example. This is largely due to the fact that normal instances are too close to anomalies. Supervised classification models based on ensembles of decision trees coped with the classification task perfectly, showing approximately 99% of correct answers, high speed of training and operation.

Conclusion: the experience gained can be used to detect accidents in various areas of human activity. In future work, it is planned to study the capabilities of classification algorithms to determine specific classes of events (pre-accident and emergency) so that it is possible to recommend actions depending on the class of malfunction. Very often, the data is temporal, and the time sign plays a significant role in predicting emergency situations using regression algorithms, which must also be taken into account to improve the results.

РЕЗЮМЕ

А. В. Звягинцева, И. Ю. Ковалев

Анализ моделей классификации для распознавания прецедентных событий в технологических процессах добычи нефти и газа

Нефтегазовая отрасль играет важнейшую роль в мировой экономике, оказывая влияние на транспорт, производство электроэнергии, нефтехимическую промышленность и национальную безопасность. Аварии, такие как взрывы на платформах или разливы нефти, приводят к серьезным экологическим катастрофам и человеческим жертвам. Своевременное распознавание опасных событий на основе данных технологического мониторинга, а также выявление прецедентов по ретроспективным данным – актуальная задача для многих отраслей промышленности. Цель работы – провести сравнительный анализ моделей классификации для выявления прецедентов на примере датасета аварий на нефтяных скважинах с естественным притоком нефти и газа.

Выполнен анализ и обработка данных, проведена настройка гиперпараметров каждой модели при помощи инструментов GridSearchCV и RandomizedSearchCV, выполнено обучение моделей двух категорий: без учителя (One Class SVM, Local outlier factor, Isolation Forest), с учителем (Linear SVM, RBF SVM, Naive Bayes, Neural Net, QDA, AdaBoost, Nearest Neighbors, CatBoost, Histogram-Based Gradient Boosting, Decision Tree, XGBoost, Extra Trees, Random Forest, Bagging Classifier), проведено тестирование и рассчитаны метрики *Precision* (точность), *Recall* (полнота), *F1 (F-мера)*, *ROC AUC*, время обучения, время тестирования.

Модели классификации без учителя в теории хорошо подходят для выявления аномалий в неразмеченных данных, однако на практике для классификации прецедентных событий эти модели на примере датасета аварий показали плохие результаты (40–58% правильных ответов). В большей степени это связано с тем, что нормальные экземпляры находятся слишком близко к аномалиям. Модели классификации с учителем на основе ансамблей деревьев решений справились с задачей отлично, показав примерно 99% правильных ответов, высокую скорость обучения и работы.

Полученный опыт можно использовать для обнаружения аварий в различных сферах деятельности человека. В дальнейшей работе планируется исследовать возможности алгоритмов классификации для определения конкретных классов событий (предаварийных и аварийных), чтобы можно было рекомендовать действия в зависимости от класса неисправности. Очень часто данные являются темпоральными, и признак времени играет значительную при прогнозирования аварийных ситуаций с помощью алгоритмов регрессии, что также необходимо учесть для улучшения результатов.

Звягинцева Анна Викторовна – д.т.н., доцент, профессор кафедры компьютерных технологий ФГБОУ ВО «ДонГУ», 283001, Донецк, ул. Университетская, 24, zviagintsevaav@gmail.com. *Область научных интересов*: системный анализ, событийная и комплексная оценка; безопасность и управление социально-экономическими и техногенными системами; информационно-аналитические системы; обработка и анализ данных. Число научных публикаций – более 150.

Ковалев Илья Юрьевич – аспирант кафедры компьютерных технологий ФГБОУ ВО «ДонГУ», 283001, Донецк, ул. Университетская, 24, ilyakovalev2023@mail.ru. *Область научных интересов*: машинное обучение, рекомендательные системы, нейронные сети, интеллектуальный анализ данных, прогнозирование событий. Число научных публикаций – более 3.

Статья поступила в редакцию 01.06.2025