Санкт-Петербургский государственный университет

Королихин Владимир Игоревич

Выпускная квалификационная работа

Разработка автоматизированной системы определения операций бурения на основе машинного обучения

Уровень образования: бакалавриат

Направление *09.03.04 «Программная инженерия»*

Основная образовательная программа *СВ.5080.2017 «Программная инженерия»*

Научный руководитель:
к.т.н., доцент М.А. Серов

Рецензент:
Главный специалист ООО "Системы Компьютерного Зрения" Д.Н. Степанов

Санкт-Петербург
2021

Saint Petersburg State University

Vladimir Korolikhin

Bachelor's Thesis

Development of an automated system for determining drilling operations based on machine learning

Education level: bachelor

Speciality *09.03.04 "Software Engineering"*

Programme *CB.5080.2017 "Software Engineering"*

Scientific supervisor:
C.Sc., docent M.A. Serov

Reviewer:
Chief Specialist at LLC Computer Vision Systems D.N. Stepanov

Saint Petersburg
2021

Оглавление

Введение

Одной из основных задач нефтегазовых компаний на сегодняшний день является цифровая трансформация отрасли в целом и функции бурения в частности. Ввиду большого количества поступающей информации с месторождений и постоянного роста числа регистрируемых параметров, оказались востребованными системы, интерпретирующие комплекс параметров и выдающие специалисту промежуточный результат в виде рекомендаций. Это облегчает анализ текущей ситуации и повышает оперативность принятия решений.

Одной из задач, решаемых такими системами, является определение операций в процессе бурения скважины. Такие операции естественным образом возникают при проведении любого вида буровых работ. Они характеризуют различные виды спусков, подъемов, вращения инструментов и многое другое, что впоследствии является основой для составления оптимизационных планов при строительстве скважин. Определяются операции на основе геолого-технической информации, поступающей с буровых площадок, однако часто оказывается, что они обладают специфической структурой, которую сложно в точности воспроизвести алгоритмически. Несмотря на это, учитывая особенности входных данных при разработке алгоритмов, можно значительно повысить их точность и эффективность.

Геолого-технические параметры для контроля строительства скважины поступают в центр управления бурением, где обрабатываются и отображаются в режиме реального времени на экранах, установленных в офисе сотрудника компании. Количество наблюдаемых параметров может варьироваться в зависимости от конкретной операции бурения, а также от типа исследования. Например, геофизические исследования направлены на изучение геологического разреза скважин. Благодаря им можно получить информацию о свойствах окружающих пород. Другой тип исследования заключается в анализе состояния бурового долота: вес на крюке, скорость проходки и прочее. Чем больше имеется информации о текущем состоянии бурения, тем больше закономерно-

стей можно найти при определении той или иной буровой операции.

Существенной проблемой при работе с реальными данными является возможное присутствие различного рода шумов, что делает точные методы не применимыми. Распространенный способ обработки таких данных — использование методов машинного обучения, в особенности нейронных сетей. Кроме того, нейронные сети предоставляют возможность эффективно находить сложные и не поддающиеся формализации структурные закономерности входных данных.

В данной работе рассматриваются различные подходы классификации операций бурения по данным геолого-технической информации. Предложенные алгоритмы реализованы в виде отдельного программного модуля для удобного пользования в промышленных компаниях. Он позволяет облегчить мониторинг бурения скважин, а также снижает количество ошибок классификации за счет использования дополнительных параметров.

1. Постановка задачи

Целью данной работы является разработка автоматизированной системы классификации операций бурения на основе геолого-технических данных с использованием методов машинного обучения. Для достижения цели были поставлены следующие задачи:

1. Сделать обзор предметной области.
2. Реализовать выбранные алгоритмы машинного обучения и сравнить их по метрикам качества.
3. Провести эксперименты и проанализировать результаты.
4. Разработать прототип веб-сервиса.

2. Обзор

В этом разделе рассматривается процесс бурения нефтяных скважин, набор данных, который поступает с датчиков на буровой площадке и основные буровые операции, используемые в нефтедобывающей промышленности. Также производится обзор алгоритмов, решающих задачу классифиции буровых операций.

2.1. Процесс бурения нефтяных скважин

Для строительства высокотехнологичных скважин нефтяные компании используют большое количество современного оборудования: буровые вышки, колонны, долота и другое. При этом строительство должно находиться под постоянным контролем специалистов. В их распоряжении находятся системы каротажа, которые используются для проведения детальных геофизических исследований. На их основе принимаются решения о текущей обстановке на буровой площадке, проводятся корректировки бурового долота в нужном направлении, фиксируются сбои, а также многое другое, без чего добыча нефти была бы невозможной.

Основополагающим способом изучения и анализа обстановки на буровой площадке является использование станции геолого-технологических исследований скважин (ГТИ), которая в режиме реального времени фиксирует процессы, происходящие на площадке. Количество замеряемых характеристик, их объем, единицы измерения могут отличаться в зависимости от места бурения. Нередки случаи, когда станция ГТИ замеряет не все заявленные характеристики, что влечет за собой пропуски в данных. Такие ситуации требуют дополнительного внимания человека, который должен самостоятельно анализировать обстановку и примать решения.

Определение операций бурения является важной задачей, так как эти операции включают в себя все процессы, происходящие во время работы со скважиной. Ошибочное определение может повлечь серьезные финансовые и временные затраты.

Основным способом для анализа данных со станции ГТИ является

сравнение каждой характеристики по заранее определенной методике. В зависимости от значения того или иного параметра делается предположение о том, какая операция происходит на данный момент. Недостатком такой методики является ее неоднозначность, так как сравнение с пороговыми значениями сильно огрубляет исходные данные. Кроме того, окончательные результаты базируются на эмпирических знаниях о предметной области сотрудника компании, который контролирует правильность определения операции. Таким образом, возникает необходимость автоматизации определения операций бурения.

2.2. Набор данных станции ГТИ

Станция ГТИ определяет целый комплекс параметров, который потенциально можно использовать для определения операций бурения. Он может включать физические свойства горных пород, содержание и состав газов в буровом растворе, описывать тектоническую обстановку и многое другое. Однако, в данной работе мы остановимся лишь на тех, которые характерезуют состояние буровых инструментов в фиксированные моменты времени. Стоит отметить, что такой комплекс параметров используется в методике промышленных нефтедобывающих компаний, а также является постоянным на каждой буровой площадке.

Согласно исследованию, проведенному в [4] основными техническими параметрами, поступающими в центр управления бурением, являются: глубина скважины, вертикальная глубина, скорость проходки, высота крюка, вес на долоте, нагрузка на долото, вертикальное смещение буровой установки.

На основе данных параметров можно делать вывод о том, как идет ход бурения, а потому именно эти параметры используется в алгоритмах, которые рассмотрены ниже.

2.3. Виды операций бурения

Процесс разработки нефтяной скважины не является непрерывным процессом, состоящим из одной операции. Он представляет последователь-

ность выполнения операций бурения, которыми, например, могут быть спуск или подъем. Во время этих операций буровое долото смещается и настраивается на нужную глубину. Также существует непосредственно сам процесс бурения. Его можно разбить на 6 основных буровых операций [4]: Роторное бурение, Роторное расширение, Скользящее бурение, Скользящее бурение Регулировка инструмента или обратное бурение, Вскрытие, Вращение.

2.4. Обзор алгоритмов

Существующие алгоритмы в общем случае классифицируют буровые операции, рассмотренные выше. Для извлечения нетривиальной, ранее неизвестной и потенциально полезной информации лучше всего подходят алгоритмы машинного обучения. Рассмотрим их подробнее в следующих разделах.

2.4.1. Муравьиный алгоритм

Рассмотрим подход, описанный в статье [8]. В этой работе исследуется гибридный алгоритм на основе метода роя частиц и муравьиного алгоритма (PSO/ACO) для классификации операций бурения скважин. Алгоритм PSO/ACO обнаруживает правила классификации IF-THEN из данных с использованием непрерывных и категориальных признаков без преобразования категориальных значений в числа.

PSO/ACO был придуман и описан Холденом и Фрейтасом в [5]. Искомые правила состоят из антецедента (набора значений признаков) и консеквента (класса). Их можно выразить следующим образом:

$$IF\langle attrib = value\rangle AND ... AND\langle attrib \geq value\rangle THEN\langle class\rangle$$

Классификатор на основе PSO/ACO был обучен и протестирован на данных по месторождению в Бразилии. На обучающей выборке точность составила 92.53%, на тестовой 94.08%. Стоит отметить, что PSO/ACO со 100% точностью определил операции вращения и отключения.

2.4.2. Метод опорных векторов

Метод опорных векторов является одним из самых популярных методов машинного обучения. Основная идея метода заключается в построении гиперплоскости, разделяющей объекты выборки оптимальным способом.

В статье [4] применяется данный подход для классификации буровых операций. В самом общем случае метод опорных векторов решает задачу бинарной классификации. Однако буровых операций гораздо больше, поэтому необходима модификация, расширяющая его на большее количество классов. Задачу многоклассовой классификации можно свести к решению нескольких бинарных задач. Существует две основные стратегии: «каждый против каждого» и «один против всех». В статье [6] сравниваются оба метода. Было показано, что модели на основе стратегии «каждый против каждого» обучаются быстрее, так как разбивают задачу на множество более мелких. Кроме того, в случае классификации буровых операций нет гарантии, что существуют хорошие различия между одним классом и остальными.

Используя метод опорных векторов, авторам статьи [4] удалось добиться 93.73% точности на обучающей и 92.6% на тестовой выборках. Авторы отмечают, что худшие результаты были получены для роторного бурения и роторного расширения. Это связано с тем, что характеристики операций сильно похожи.

2.4.3. Нейронная сеть типа MPL

При разработке системы автоматической классификации буровых операций авторы статьи [3] использовали нейронную сеть типа Perceptron Multiple Layers (MPL) с 5 нейронами во входном слое представляющих параметры каротажа, 5 нейронов в среднем слое и 1 нейрон в выходном, представляющий один из возможных этапов операции бурения (см. рис. 1).

Количество нейронов в среднем слое было выбрано экспериментально после тестирования модели, которая повысит процент верных пред-

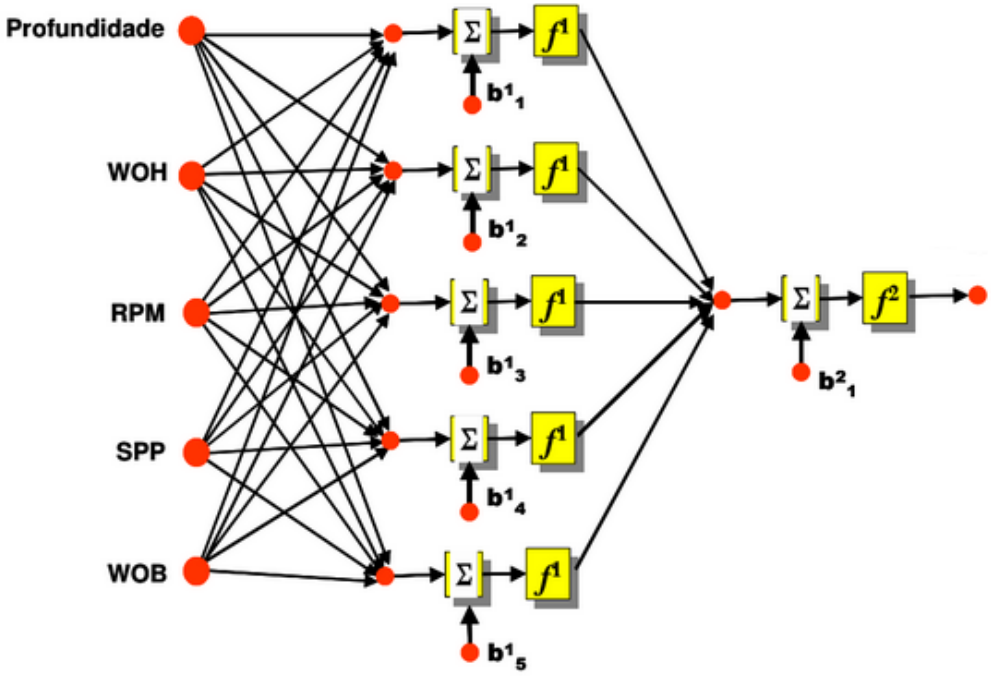

Рис. 1: Архитектура нейронной сети, предложенная в статье [3]

сказаний. В этом выборе также учитывался компромисс между временем обучения и временем предсказания на тестовой выборке, так как данную систему планировалось использовать для классификации в реальном времени.

Предложенная модель была обучена и протестирована на реальных данных, которые были получены при бурении морской скважины, расположенной на нефтяном месторождении в Бразилии. На тестовом наборе точность составила 94%, что говорит об эффективности модели.

2.4.4. Одномерная свёрточная нейронная сеть

Одномерная свёрточная нейронная сеть использовалась для классификации литологических фаций [7]. Фациии обозначают свойства осадочных горных пород и представляют интерес для исследователей.

Архитектура сети изображена на рис. 2. Она состоит из входного слоя (7-мерный вектор), четырех сверточных слоев с функцией активации ReLU, двух max-pooling слоев и трех полносвязных слоев. Accuracy на тестовых данных составила 77%.

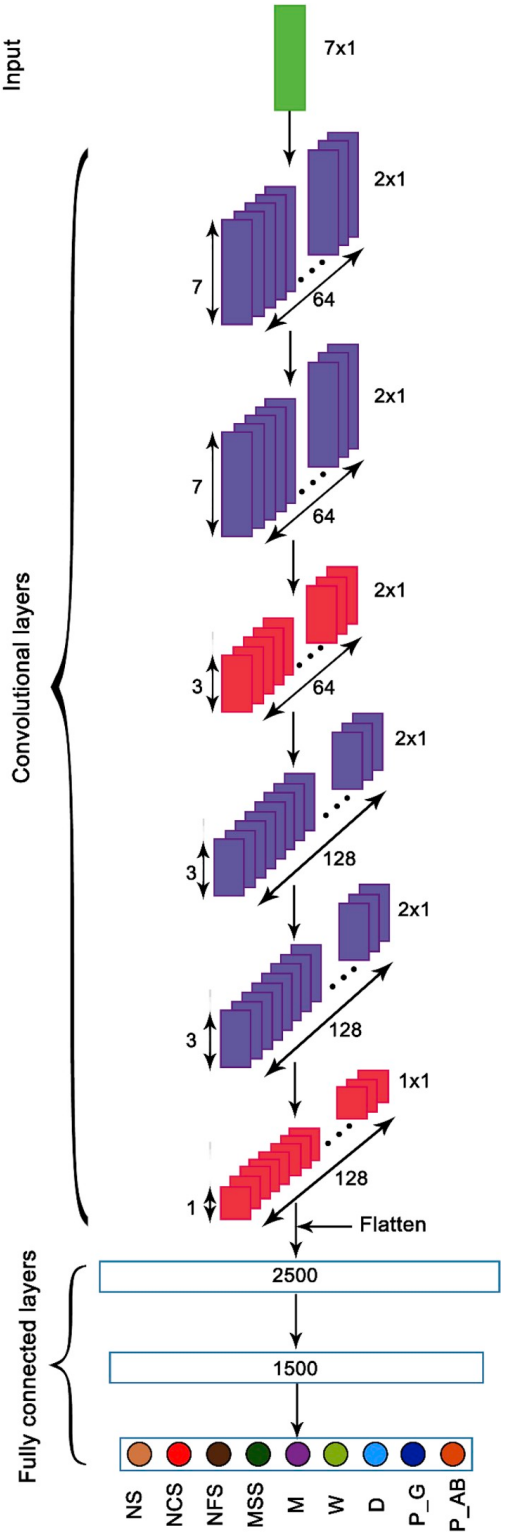

Рис. 2: Архитектура одномерной сверточной нейронной сети для классификации фаций из статьи [7]

2.5. Вывод

Существующие подходы ориентированы в основном на анализ 6 буровых операций и используют в качестве признаков геофизические параметры и показатели работы приборов. Лучшие результаты были получены в работе [3] с использованием нейронных сетей. Также стоит отметить высокие показатели точности муравьиного алгоритма и метода опорных векторов.

3. Задача классификации буровых операций

В этом разделе описаны данные, которые используются для классификации операций. Рассмотрен метод их нормализации, а также генерации признаков. Приведены апробированные методы машинного обучения и результаты их сравнения по метрикам качества.

3.1. Описание исходных данных

Данные с буровых площадок были предоставлены компанией ПАО «Газпром нефть» [10]. Газпром нефть является одним из лидеров российского нефтегазового рынка по объемам добычи, во многом за счет строительства высокотехнологичных скважин, а также благодаря методам обработки поступающей информации.

Областью исследований является Новопортовское нефтегазовое месторождение на полуострове Ямал [9]. Всего было произведено 5000 замеров (т.е. 5000 векторов измеренных признаков), взятых с интервалом по времени в 1 секунду. Каждый набор был помечен типом операции бурения. Количество операций бурения (классов) в текущем исследовании равняется 22. Кроме того, само бурение составляет около 20% от общего числа операций на скважине.

Вектор признаков содержит в себе большое количество геофизических показателей среды, а также различные характеристики нагрузки на приборы, однако наиболее универсальными и информативными являются последние. В данной работе мы использовали следующий набор признаков: Глубина долота, Глубина забоя, Нагрузка на долото, Обороты ротора/ВСП, Расход на входе, Расход на выходе, Давление на входе, Плотность на входе, Момент на ключе и Давление на входе.

В Таблице 1 приводится список, рассматриваемых нами операций бурения. Заметим, что в данных присутствует несколько видов "Спуска", "Подъема", "Промывки" и "Бурения", поэтому такие операции можно разбить на группы: Спуск, Подъем, Промывка, Бурение. Остальные

Группа	Операция
Спуск	Спуск с промывкой Спуск с проработкой
	Спуск в скважину
	Спуск бурильной свечи
	Спуск свечи с вращением и циркуляцией на длину свечи
	Спуск обсадной колонны на длину трубы в скважину
	Допуск бурильной колонны до забоя
Подъем	Подъем с промывкой
	Подъем с проработкой
	Подъем из скважины
	Подъем свечи с вращением и циркуляцией на длину свечи
	Подъем бурильной колонны с вращениеми циркуляцией на длину трубы
	Подъем бурильной колонны на длину свечи
Промывка	Промывка с вращением
	Промывка при неподвижном состоянии
Бурение	Роторное бурение
	Выработка нагрузки (Роторное бурение)
	Направленное бурение
	Выработка нагрузки (Направленное бурение)
Наращивание	Наращивание
Неподвижное состояние	Неподвижное состояние

Таблица 1: Список операций бурения, определяемых в данной работе

же операции будем определять по одной, так как их нельзя отнести ни к одной из введенных групп, а также объединить между собой. Кроме того, в имеющихся данных наблюдается дисбаланс классов. Преобладает спуск, а наращивание представлено очень скудно (см. рис. 3).

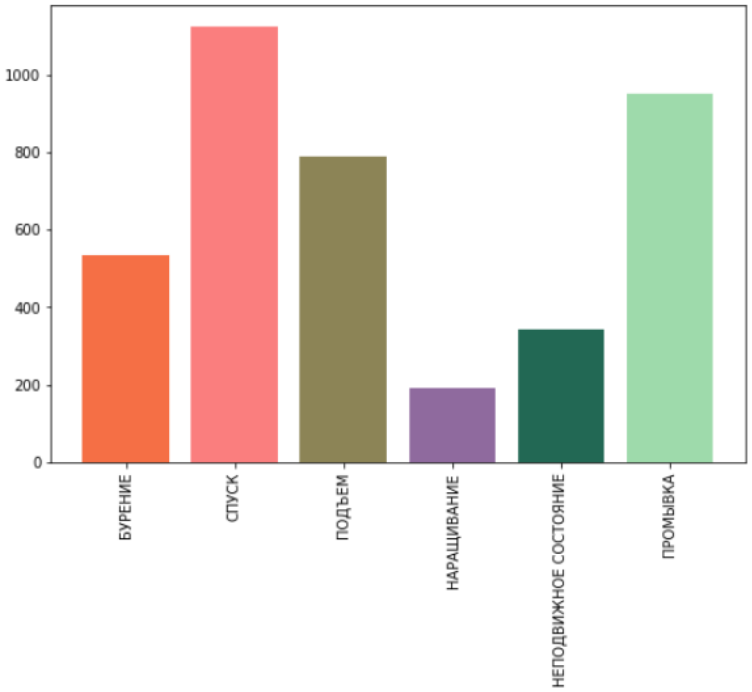

Рис. 3: Число объектов для каждой из групп

На основании признакового описания объектов, а также пользуясь методикой, предоставленной компанией Газпром нефть, можно эмпирическим путем определить некоторые операции. На Спуск и Подъем оказывает большое влияние изменение глубины долота и забоя, нагрузка на долото и обороты ротора. При бурении и промывке показатели оборотов ротора должны быть больше нуля. В случае проработки следует ориентироваться на показатели нагрузки на долото и скорости инструмента.

На практике классификатор, построенный по правилам простого ветвления, то есть сравнения каждого признака с некоторым пороговым значением часто ошибается, а именно точность классификации составляет не больше 80%. Кроме того, из-за большого количества операций бурения, а также количества признаков их описывающих такой классификатор представляется большой ветвящейся структурой, слож-

но масштабируемой для новых признаков. Таким образом, применение машинного обучения для классификации операций бурения является оправданным решением.

3.2. Предобработка исходных данных

Качество классификации с помощью алгоритмов машинного обучения сильно зависит от многих факторов, в частности от качества исходных данных. Несмотря на это, во многих предметных областях приходится сталкиваться с данными, которые имеют тенденцию быть неполными. Они могут содержать шумы, а также иметь большой разброс значений. На таких данных многие алгоритмы машинного обучения могут вести себя не так, как ожидалось. Следовательно, данные нуждаются в предварительной обработке.

В нашей задаче классификации операций бурения, признаковое описание объектов было получено на основе фактической работы приборов. Поэтому влияние различного рода шумов можно считать минимальным. Кроме того, в связи с тем, что буровые площадки были выбраны таким образом, чтобы в них содержались все признаки, описанные ранее, решать задачу пропущенных значений нам не потребовалось. Однако, в данных есть другой вид "несовершенства", от которого нужно избавиться. Необходимо устранить сильное различие между значениями признаков и придать им распределение с меньшей амплитудой. В рамках данного исследования были рассмотрены несколько подходов для дальнейшего использования в методах машинного обучения:

1. Минимаксная нормализация

   Каждый признак x_i преобразуется к $x_i^{'} = \frac{x_i - x_{min}}{x_{max} - x_{min}}$. Здесь x_{min} - минимальное значение признака, а x_{max} - максимальное.

2. Нормализация средним

   Каждый признак x_i преобразуется к $x_i^{'} = \frac{x_i - \mu}{\sigma}$, где $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$ является средним значением признака x, а $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$ это его стандартное отклонение.

Такое преобразование приводит каждый признак выборки к величине со стандартным нормальным распределением.

3. Квантильное преобразование

   Каждый признак x_i вычисляется с помощью нелинейного преобразования:
   $x_i^{'} = \Phi_i^{'}(F(x_i))$ Здесь Φ — функция распределения стандартного нормального закона, F — эмпирическая функция распределения.

   В его результате признак будет принимать значения из [0, 1].

4. Преобразование Йео-Джонсона

   Преобразование Йео-Джонсона предназначено для стабилизации дисперсии и минимизации асимметрии выборки. Оптимальный параметр λ, участвующий в преобразовании, оценивается для каждого признака отдельно, путем минимизации функции максимального правдоподобия.

   В данном преобразовании каждый признак x_i вычисляется следующим образом:

$$x_i^{'} = \begin{cases} \frac{(x_i+1)^\lambda}{\lambda}, & \text{если } \lambda \neq 0, x_i \geq 0 \\ \log(x_i+1), & \text{если } \lambda = 0, x_i \geq 0 \\ \frac{-[(x_i+1)^{(2-\lambda)}-1]}{2-\lambda}, & \text{если } \lambda \neq 2, x_i < 0 \\ -\log(-x_i+1), & \text{если } \lambda = 2, x_i < 0 \end{cases}, \text{где } 0 \leq \lambda \leq 2$$

3.3. Метрики качества

Данная задача является одной из классических задач в машинном обучении, а именно задачей многоклассовой классификации. В качестве метрики для оценки результата работы алгоритмов классификации операций бурения были выбраны accuracy, precision, recall и f1-мера. Опишем их подробнее:

1. Accuracy (доля правильных ответов) вычисляется как $\frac{TP+TN}{TP+TN+FP+FN}$. Преимуществом такой метрики является ее интуитивность. Недостатком же является проблема с интерпретацией на несбалансированных выборках.

2. Precision (точность) показывает насколько можно доверять классификатору и вычисляется как $\frac{TP}{TP+FP}$.

3. Recall (полнота) показывает как много объектов находит классификатор. Вычисляется по формуле $\frac{TP}{TP+FN}$

4. F1-мера позволяет найти некий баланс между Precision и Recall, объединяя в себе значения этих метрик. Вычисляется по формуле $2 * \frac{Precision*Recall}{Precision+Recall}$.

Кроме того, решая задачу многоклассовой классификации, необходимо иметь ввиду, что нам нужно каким-то образом усреднять результат для каждой из метрик. Рассмотрим два возможных похода:

- Микро-усреднение. Используется подход "один против всех": строится количество бинарных классификаторов, равное количеству классов в выборке. Для каждой из задач вычисляются TP, TN, FP и FN. Результирующие метрики находятся по формулам:

  $Recall_{micro} = \frac{\sum_i TP_i}{\sum_i TP_i + \sum_i FP_i}$ $\qquad$ $Precision_{micro} = \frac{\sum_i TP_i}{\sum_i TP_i + \sum_i FN_i}$

  Стоит заметить, что при микро-усреднении результирующие метрики зависят от размера каждого из классов.

- Макро-усреднение. Аналогично микро-усреднению здесь используется подход «один против всех», однако результирующие метрики, являются усреднением метрик по каждому из классов (n штук):

  $Recall_{macro} = \frac{\sum_i^n Recall_i}{n}$ $\qquad$ $Precision_{macro} = \frac{\sum_i^n Precision_i}{n}$

  В данном подходе каждый класс вносит одинаковый вклад в результирующие метрики.

В нашей работе мы будем пользоваться макро-усреднением для оценки f1-меры, потому что размеры классов различаются довольно сильно.

3.4. Методология

Для решения поставленной задачи многоклассовой классификации операций бурения была выбрана следующая методология: предварительно классифицировать на 6 буровых групп (Спуск, Подъем, Промывка, Бурение, Наращивание и Неподвижное состояние), после этого классифицировать операцию пользуясь знанием о ее группе.

Для классификации на группы были выбраны следующие методы машинного обучения:

1. Метод опорных векторов
2. Градиентный бустинг над решающими деревьями
3. Логистическая регрессия
4. Нейронная сеть типа MPL
5. Одномерная свёрточная нейронная сеть

Метод опорных векторов и нейронная сеть хорошо показали себя на этапе обзора. Градиентный бустинг над решающими деревьями был использован в работе классификации фаций по данным ГТИ и показал лучшие результаты [1] среди рассматриваемых методов. Логистическая регрессия взята по причине хорошей интерпретируемости своих результатов.

Обзор каждого из алгоритмов будет приведен ниже. Для всех этих алгоритмов настроены параметры, а также описана архитектура нейронной сети. Эти алгоритмы, а также их результаты были получены при классификации на 6 буровых групп.

3.4.1. Метод опорных векторов

Данный метод заключается в поиске разделяющей гиперплоскости с наибольшим зазором между классами. При этом исходные признаковые

описания объектов переводятся в пространство большей размерности. Классификатор при обучении старается минимизировать среднеквадратичную ошибку, которая тем меньше, чем больше расстояние между двумя параллельными гиперплоскостями, построенными обеим сторонам гиперплоскости, разделяющей классы.

Для улучшения метрик качества многоклассовой классификации методом опорных вектором были выбраны следующие параметры для настройки: параметр регуляризации (C), тип ядра (kernel), преобразующий вектор признаков в пространство другой размерности (возможно большей), чтобы добиться линейной разделимости.

Лучшая комбинация параметров была найдена с помощью поиска по сетке и кроссвалидации по 5 блокам. При этом C увеличивалось в 10 раз, начиная с 1 и закачивая 1000. Использовались две функции ядер: linear и rbf. В случае rbf ядра также настраивался параметр gamma на значениях от 0.1 до 0.0001, увеличивая в 0.1 на каждой итерации.

Метрики / Преобразование	Accuracy	$F1 - score$
Минимаксная нормализация	43.45	52.17
Нормализация средним	47.45	51.74
Квантильное преобразование	76.61	70.77
Преобразование Йео-Джонсона	52.55	53.81

Таблица 2: Значения метрик качества классификации объектов на группы

Самое высокое значение accuracy составляет 76% в случае квантильного преобразования при выборе rbf ядра с параметрами $C = 1000$ и $gamma = 0.01$. При этом f1-мера составила 71%. Сводная информация по остальным видам преобразований представлена в Таблице 2.

3.4.2. Градиентный бустинг над решающими деревьями

Данный метод строит композицию деревьев решений и вычисляет итоговую вероятность принадлежности классу как взвешенную сумму ве-

роятностей для каждого из классификаторов. Основная идея заключается в том, что модель учится на своих же ошибках, стараясь их минимизировать на следующих итерациях. В качестве функции потерь используется среднеквадратичное отклонение (MSE).

Существует множество вариаций алгоритмов градиентного бустинга над решающими деревьями. Одним из самых эффективных является экстремальный градиентный бустинг (XGBoost) [2].

Для улучшения метрик качества многоклассовой классификации методом XGBoost были выбраны следующие параметры для настройки: количество деревьев решений (n_estimators), ограничение на их максимальную глубину (max_depth), а также скорость обучения (learning_rate). Параметр максимальной глубины нуждается в особом внимании, так как слишком глубокие деревья могут запомнить информацию, присущую именно обучающей выборке, а недостаточная глубина наоборот, не позволяет захватить всех особенности в данных.

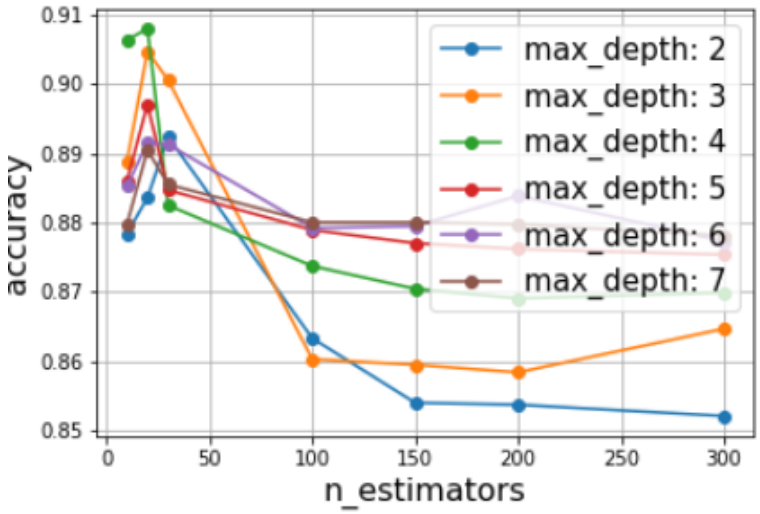

Рис. 4: Зависимость accuracy от количества деревьев решений

Лучшая комбинация параметров была найдена с помощью поиска по сетке и кроссвалидации по 5 блокам. При этом n_estimators увеличивалось на 50, начиная от 100 и заканчивая 450, max_depth в диапазоне от 3 до 15 с шагом 1, а learning_rate с 0.1 до 1 с шагом 0.1.

Метрики / Преобразование	Accuracy	$F1-score$
Минимаксная нормализация	88.55	87.98
Нормализация средним	87.54	86.92
Квантильное преобразование	89.58	88.84
Преобразование Йео-Джонсона	90.48	89.80

Таблица 3: Значения метрик качества классификации объектов на группы

Самое высокое значение accuracy составляет 91% в случае преобразования Йео-Джонсона при $learning\_rate = 0.2$, $max\_depth = 6$ и $n\_estimators = 100$. При этом f1-мера равна 90%. На рис. 4 изображен график зависимости метрики accuracy от количества деревьев решений при фиксированном $learning\_rate = 0.2$. Сводная информация по остальным видам преобразований представлена в Таблице 3.

3.4.3. Логистическая регрессия

Логистическая регрессия является одним из статистических методов классификации. Она оценивает вероятность того, что вектор признаков принадлежит каждому классу. Преимуществом такого подхода является высокая интерпретируемость: для каждого признака можно взять его коэффициент и оценить степень важности для каждого класса, тем самым получая дополнительную информацию о классах.

Основным параметром для настройки является параметр регуляризации C, а также методы регуляризации. Лучшая комбинация параметров была найдена с помощью поиска по сетке и кроссвалидации по 5 блокам. При этом C увеличивалось в 10 раз, начиная с 1 и заканчивая 1000. В качестве методов регуляризации были взяты L_1 и L_2.

Результаты экспериментов приведены в Таблице 4.

Преобразование \ Метрики	Accuracy	$F1 - score$
Минимаксная нормализация	38.42	43.35
Нормализация средним	47.45	51.74
Квантильное преобразование	78.42	79.04
Преобразование Йео-Джонсона	52.36	58.57

Таблица 4: Значения метрик качества классификации объектов на группы

3.4.4. Нейронная сеть типа MPL

Нейронные сети являются мощным инструментом в решении задач классификации в тех областях, где данные сложно формализуются. Архитектура нейронной сети уникальна для каждой задачи. Однако, на основе проведенного обзора, а также эксперимента с различными слоями была выбрана архитектура, изображенная на рис. 5. Наиболее высокую эффективность показало чередование полносвязных слоев (Dense) и Dropout слоев с нормализацией, которая использовалась для стабилизации процесса обучения.

Результаты экспериментов приведены в Таблице 5.

Преобразование \ Метрики	Accuracy	$F1 - score$
Минимаксная нормализация	39.64	46.47
Нормализация средним	51.88	53.88
Квантильное преобразование	83.42	82.03
Преобразование Йео-Джонсона	43.34	42.55

Таблица 5: Значения метрик качества классификации объектов на группы

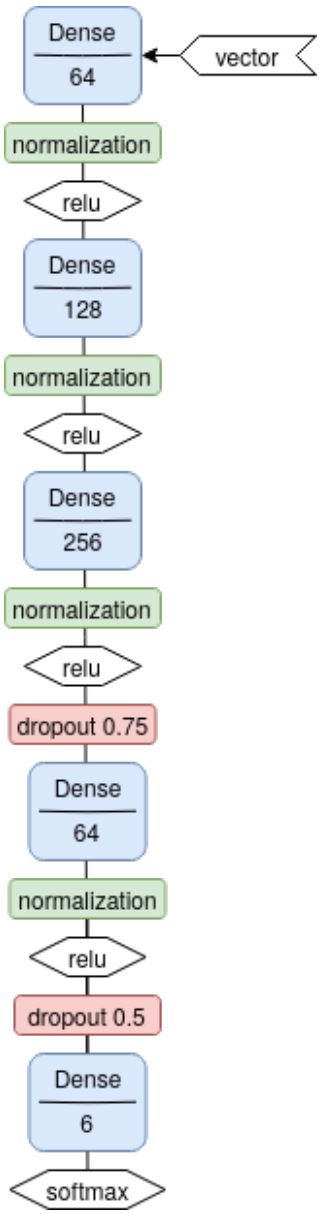

Рис. 5: Архитектура нейронной сети типа MPL для классификации на группы

3.4.5. Одномерная свёрточная нейронная сеть

Рассмотренная в обзоре модель была реализована с гиперпараметрами, которые были описаны в статье [7]. Несмотря на это, в ходе экспериментов было установлено, что batch_size = 10 с количеством эпох равным 1000 являются более оптимальными (в отличие от 4000 эпох) с точки зрения эффективности и точности модели.

Результаты классификации отражены в Таблице 6.

Преобразование \ Метрики	Accuracy	$F1 - score$
Минимаксная нормализация	40.60	47.47
Нормализация средним	52.45	54.45
Квантильное преобразование	84.23	82.53
Преобразование Йео-Джонсона	45.35	43.66

Таблица 6: Значения метрик качества классификации объектов на группы

3.5. Предлагаемое решение

Настроенные на предыдущем этапе модели машинного обучения для решения задачи классификации групп операций бурения были использованы в качестве первого этапа классификации. Для определения самих операций использовалась информация о группах, закодированная с помощью унитарного кода. Таким образом, список признаков пополнился еще 6 значениями.

На предыдущих этапах наибольшее значение метрик accuracy и f1-меры удалось достичь с помощью преобразования Йео-Джонсона методом градиентного бустинга над решающими деревьями. Поэтому для классификации результирующих операций был выбран также этот метод. Итоговые метрики качества (с учетом потерь на этапе определения групп) составили: $accuracy = 85\%, f1 - score = 74\%$

Стоит отметить, что классификатор не смог определить следующие операции: Выработка нагрузки (направленное бурение), спуск свечи с вращением и циркуляцией на длину трубы и спуск обсадной колонны на длину трубы в скважину. Это связано с тем, что такие операции встречаются очень редко, а потому составляют всего 2% на каждый класс от общего количества тренировочных данных.

4. Эксперименты

4.1. Классификация групп: спуск, подъем, бурение и промывка

На группы спуск, подъем, бурение и промывка приходится большая часть времени, затраченного при разработке месторождения. Неподвижное состояние и наращивание встречаются реже остальных, поэтому в работе рассмотрена возможность классификации отдельно на данные группы.

Результаты классификации при сокращении количества групп значительно улучшились. Наибольшие показатели по прежнему у преобразования Йео-Джонсона. Оценки по метрикам качества приведены в таблице 7

Метрики / Преобразование	Accuracy	$F1 - score$
Минимаксная нормализация	94.22	93.54
Нормализация средним	94.34	93.23
Квантильное преобразование	95.23	94.34
Преобразование Йео-Джонсона	95.33	94.48

Таблица 7: Значения метрик качества классификации объектов на группы

4.2. Использование времени в качестве дополнительного признака

Ранее было сказано, что станция ГТИ, согласно спецификации производит замеры с буровой площадки раз в 1 секунду. Однако, на практике, это оказалось не совсем так. На рис. 6 изображена диаграмма зависимости среднего значения длительности выполнения операций в группе от самой группы.

Результаты классификации при использовании дополнительного признака приведены в таблице 8. Как видно, его использование не приве-

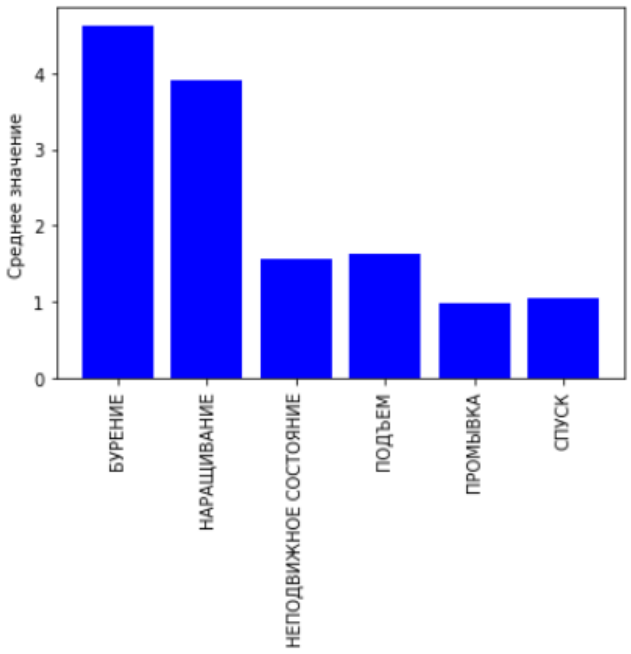

Рис. 6: Среднее значение длительности операций в группе

ло к значительным улучшениям, поэтому можно считать его влияние незначительным.

Метрики / Преобразование	Accuracy	$F1-score$
Минимаксная нормализация	89.23	88.34
Нормализация средним	86.66	87.33
Квантильное преобразование	89.33	89.45
Преобразование Йео-Джонсона	90.53	88.45

Таблица 8: Значения метрик качества классификации объектов на группы

4.3. Дискуссия и выводы

Алгоритмы машинного обучения показывают приемлемые результаты, их использование несомненно может повысить качество классификации операций бурения. Однако, стоит отметить возможные улучшения данной работы:

- Использование более репрезентативной выборки, т.е. необходимо наличие большего количества экземпляров для каждой операции.

- Увеличение интервала между замерами больше чем на 1 сек. Такое изменение позволяет лучше прослеживать дельту изменения глубины забоя и глубины долота.

В результате экспериментов было показано:

- Спуск, подъем, бурение и промывка хорошо разделимы.
- Дополнительный признак длительности по времени скорее связан с погрешностями в работе самих приборов и никак не зависит от операций.

5. Прототип веб-сервиса

В рамках работы был создан прототип веб-сервиса, предоставляющий пользователю возможность наглядно видеть список операций бурения, а также производить операции с ними. Так, система позволяет проводить фильтрацию по заданным параметрам, загружать и выгружать данные. Кроме того, рассмотренный алгоритм машинного обучения интегрирован в веб-сервис и автоматически определяет операции при загрузке в него данных.

5.1. Требования к прототипу

К веб-сервису были поставлены следующие требования:

1. Обеспечить минимальный пользовательский интерфейс в виде выпадающих списков, а также отображения информации по каждой операции в виде ее признакового описания и названия операции.

2. Обеспечить возможность загрузки/выгрузки списка операций. В выгружаемых данных должна присутствовать колонка "Операция", определяемая системой. Формат загружаемых данных .csv, выгружаемых .csv в обоих случаях.

3. Предоставить пользователю возможность запрашивать информацию выборочно по каждой скважине, а также за определенный период времени.

5.2. Выбор технологий

Для хранения данных была выбрана PostgreSQL по нескольким причинам. Во-первых данные о об операциях целостны и представляют собой структуру, которая не будет подвержена частым изменениям. К тому же в поставленной задаче не требуется обработка больших объемов данных и других специфичных условий обработки, в которых отличные от SQL хранилища были бы выгоднее. Во-вторых, PostgreeSQL является

бесплатным программным обеспечением с открытым исходным кодом: ничего не требуется платить за использование. В-третьих, PostgreeSQL охватывает большое комьюнити разработчиков, которое вносит вклад в его развитие, а также отвечает на вопросы на многочисленных онлайн-форумах.

Для разработки веб-сервиса был выбран язык программирования Python. Так как модель машинного обучения была написана на нем, то она легко интегрируется в систему.

В качестве фреймворка для веб-разработки был выбран Django. Данный инструмент является достаточно гибким и подходит практически для любого веб-сайта или приложения. Кроме того, Django имеет большое сообщество пользователей, а потому хорошо поддерживается.

Клиентская часть была написана на React по причине высокой производительности и большого сообщество разработчиков.

5.3. Архитектура и особенности реализации

Для обеспечения не только классификации операций бурения по данным ГТИ, но также и внедрения в рабочий процесс сотрудников нефтегазовых компаний, работа системы состоит из двух частей: анализ данных и взаимодействие с пользователем.

На рис. 7 представлена архитектура системы. За анализ буровых данных отвечает модуль анализа буровых данных. Данные, поступающие на вход модели, как и получающиеся в результате классификации, имеют определенный формат и размер и нуждаются в обработке. Поэтому перед классификацией и после ее завершения, данные обрабатываются компонентами предобработки входных данных и постобработки выходных соответственно.

Модель машинного обучения размещается удаленно. На это есть несколько причин. Первая заключается в сложной настройке окружения, необходимого для запуска нейронной сети. Вторая с необходимостью хранения отредактированных моделью данных. Для обеспечения удаленного доступа к модели классификации реализован контроллер

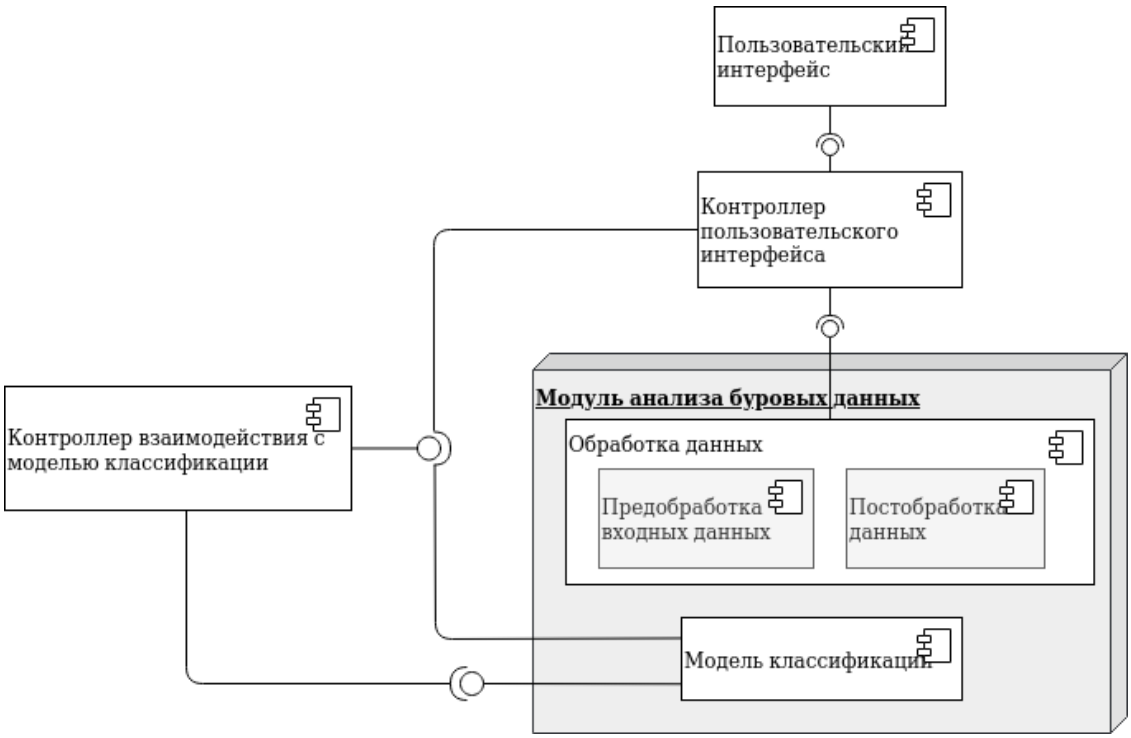

Рис. 7: Архитектура системы

взаимодействия, который представляет собой веб-сервис. Контроллер позволяет запустить модель классификации и предоставляет доступ к результатам работы модели на указанных данных для определенной скважины.

5.3.1. Модуль анализа буровых данных

Модуль классификации операций бурения состоит из двух компонентов: модели классификации и компонента, отвечающего за обработку данных.

Модель классификации осуществляет классификацию операций бурения, прошедших предобработку. Для этого используется предобученная нейронная сеть, которая принимает матрицу из данных, где каждая строка представляет набор признаков операции. Результатом работы компонента является числовой вектор, количество элементов в котором равно количеству строк в исходных данных.

Компонент обработки данных занимается обработкой входной матрицы, а также постобработкой данных, полученных в результате рабо-

ты модели классификации. Входные данные представлены в формате CSV (Comma-Separated Values). Это текстовый формат, предназначенный для представления табличных данных. После загрузки данных в систему проводится их нормализация при помощи преобразования Йео-Джонсона, которое было настроено в процессе обучения модели. Полученные в результате классификации данные переводятся в текстовое представление и сохраняются в базу данных.

5.3.2. Взаимодействие с пользователем

Для обеспечения удобного доступа к модулю классификации было принято решение реализовать пользовательский интерфейс на основе React приложения. API предоставляет возможность загрузки данных в систему, просмотр операций бурения по филиалу, скважине, дате и времени суток.

На рис. 8 изображен пользовательский интерфейс.

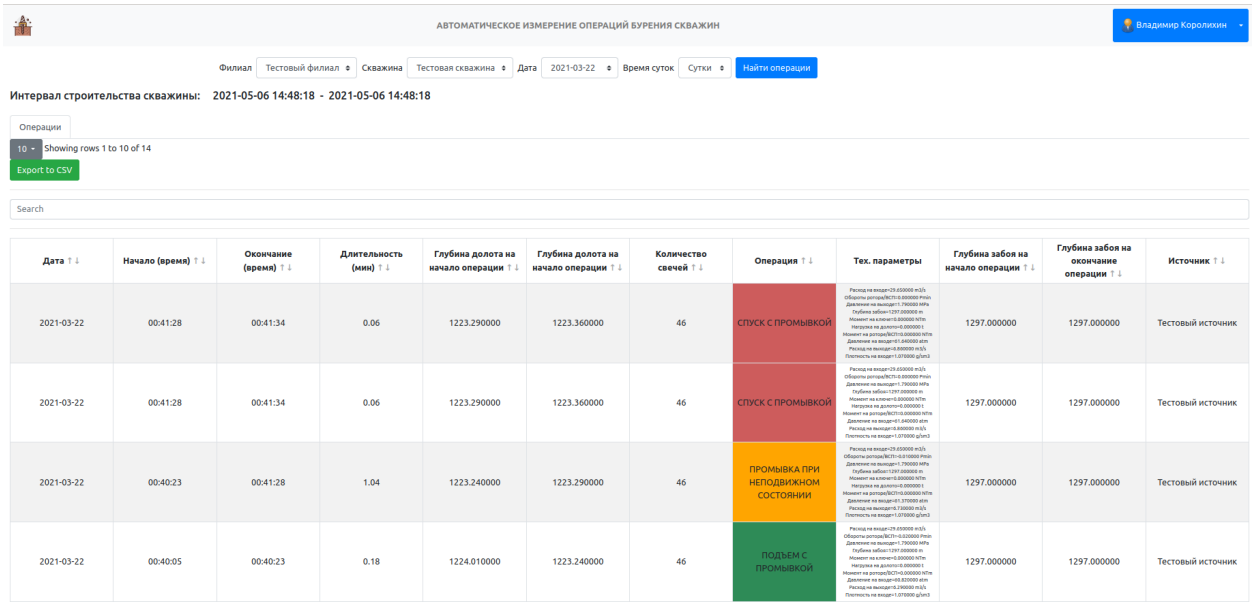

Рис. 8: Пользовательский интерфейс

6. Результаты

В ходе данной работы были получены следующие результаты.

1. Изучена предметная область. Рассмотрены основные операции, которые встречаются на буровых площадках. Также описан процесс определения буровых операций по данным каротажа.

2. Проведен анализ популярных алгоритмов машинного обучения, решающих поставленную задачу. Существующие алгоритмы классифицируют небольшое количество операций, а также используют другие геолого-технические параметры.

3. Произведена предобработка исходного набора данных с учетом выявленных особенностей. В данных удалялись ненужные признаки, вводились дополнительные и нормализовались.

4. Реализованы и сравнены алгоритмы классификации буровых операций на основе следующих алгоритмов: SVM, Logistic regression, Gradient boosting и нейронной сети. Лучше всего себя показал Gradient boosting. Accuracy на тестовых данных составила 85%.

5. Проведены эксперименты с данными: была рассмотрена возможность классификации четырех операций, наиболее часто встречающихся в процессе бурения. Также описано возможное влияние длительности операции, зафиксированное станцией ГТИ на саму операцию. Результаты показали, что такое влияние минимально.

6. Реализован прототип веб-сервиса с помощью следующего стека технологий: Django, React, PostgreSQL. Сервис классифицирует буровые операции при помощи градиентного бустинга, а также предоставляет дополнительную функциональность при работе с операциями. В частности, поддерживается возможность фильтрации операций по буровой площадке, дате, и времени суток.

Список литературы

[1] A. Hall M. Hall. Distributed collaborative prediction: Results of the machine learning contes. — 2017. — P. 267–269.

[2] Chen Tianqi, Guestrin Carlos. Xgboost: A scalable tree boosting system // Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. — 2016. — P. 785–794.

[3] Classificação automática da operação de perfuração de poços de petróleo através de redes neurais / Adriane B de S Serapião, Rogério M Tavares, José Ricardo P Mendes, Celso K Morooka // Proc. of VII Brazilian Symposium on Intelligent Automation (SBAI). São Luís-MA, Brazil. — 2005.

[4] Classification of petroleum well drilling operations using support vector machine (svm) / Adriane BS Serapiao, Rogerio M Tavares, Jose Ricardo P Mendes, Ivan R Guilherme // 2006 International Conference on Computational Inteligence for Modelling Control and Automation and International Conference on Intelligent Agents Web Technologies and International Commerce (CIMCA'06) / IEEE. — 2006. — P. 145–145.

[5] Holden Nicholas Paul, Freitas Alex A. A hybrid PSO/ACO algorithm for classification // Proceedings of the 9th annual conference companion on Genetic and evolutionary computation. — 2007. — P. 2745–2750.

[6] Hsu Chih-Wei, Lin Chih-Jen. A comparison of methods for multiclass support vector machines // IEEE transactions on Neural Networks. — 2002. — Vol. 13, no. 2. — P. 415–425.

[7] Imamverdiyev Yadigar, Sukhostat Lyudmila. Lithological facies classification using deep convolutional neural network // Journal of Petroleum Science and Engineering. — 2019. — Vol. 174. — P. 216–228.

[8] Serapião Adriane BS, Mendes José Ricardo P. Classification of petroleum well drilling operations with a hybrid particle swarm/ant colony algorithm // International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems / Springer. — 2009. — P. 301–310.

[9] Новопортовское месторождение. — Access mode: https://www.gazprom-neft.ru/company/major-projects/new-port/.

[10] ПАО «Газпром нефть». — Access mode: https://www.gazprom-neft.ru/.