

Министерство науки и высшего образования Российской Федерации  
Санкт-Петербургский политехнический университет Петра Великого  
Институт прикладной математики и механики  
Высшая школа теоретической механики

Работа допущена к защите  
Директор высшей школы  
\_\_\_\_\_ А.М. Кривцов  
«\_\_\_» \_\_\_\_\_ 2020 г.

## **ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА МАГИСТРА**

**«Разработка методологии и технологии оценки экономической  
целесообразности проведения ПГИ на основе ценности исторических  
данных о скважине»**

по направлению 01.04.03 «Механика и математическое моделирование»  
по образовательной программе  
01.04.03\_04 «Математическое моделирование процессов нефтегазодобычи»

Выполнил: студент гр.3640103/80401	_____	В. П. Казаков
Руководитель: доцент ВШТМ СПбПУ, к.т.н.	_____	П. М. Кикин
Консультант по нормоконтролю	_____	Е.А. Хайбулова

Санкт-Петербург

2020

САНКТ-ПЕТЕРБУРГСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ ПЕТРА ВЕЛИКОГО  
Высшая школа теоретической механики

УТВЕРЖДАЮ

Директор высшей школы

 А.М. Кривцов


«    » \_\_\_\_\_ 2020 г.

**ЗАДАНИЕ**

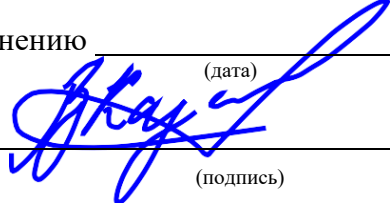
**на выполнение выпускной квалификационной работы**

студенту Казакову Василию Петровичу, группа 3640103/80401

- 1. Тема выпускной квалификационной работы:** Разработка методологии и технологии оценки экономической целесообразности проведения ПГИ на основе ценности исторических данных о скважине.
- 2. Срок сдачи студентом законченной работы:** 08.06.2020 г.
- 3. Исходные данные к работе:** Набор данных по 300 добывающим скважинам месторождения X и заключения промысловых геофизических исследований для этих скважин.
- 4. Содержание работы (перечень подлежащих разработке вопросов):** Сценарии работы добывающих скважин для реализации их в разработанных алгоритмах машинного обучения. Численные эксперименты с подбором оптимального алгоритма расчета для ожидаемой ценности проекта с учетом дополнительной информации и ожидаемой ценности проекта без учета дополнительной информации. Алгоритм общей методологии VOI для кандидатов проведения ПГИ. Реализация программного комплекса для анализа кандидатов РИР.
- 5. Перечень графического материала (с точным указанием обязательных чертежей):**
- 6. Консультант по работе:**
- 7. Дата выдачи задания:** 22.01.2020

Руководитель ВКР \_\_\_\_\_ П.М. Кикин  
  
(подпись)

Задание принял к исполнению \_\_\_\_\_  
(дата)

Студент \_\_\_\_\_ В.П. Казаков  
  
(подпись)

## **РЕФЕРАТ**

На 121 с., 11 рисунков, 3 таблицы, 2 приложения.

**ЦЕННОСТИ ИНФОРМАЦИИ, ДЕБИТ НЕФТИ, ДЕБИТ ВОДЫ,  
ПРОМЫСЛОВЫЕ ГЕОФИЗИЧЕСКИЕ ИССЛЕДОВАНИЯ, ВРЕМЯННЫЕ  
РЯДЫ, РЕМОНТНО-ИЗОЛЯЦИОННЫЕ РАБОТЫ, МАШИННОЕ  
ОБУЧЕНИЕ.**

В работе были проанализированы различные методики расчета VOI и разработан собственный метод расчета VOI, подходящий для нашей задачи. Были собраны и отсортированы данные с 30 различных добывающих скважин месторождения X. Также реализованы численные алгоритмы на полученных данных, используя методы машинного обучения и статистического анализа, на языке программирования Python. Проведена серия численных экспериментов наших предсказательных моделей на отсортированных данных с данного месторождения и подтверждена эффективность данной методики расчета VOI, с возможностью применения на других месторождениях.

## **THE ABSTRACT**

121 pages, 11 pictures, 3 tables, 2 applications.

**VALUE OF INFORMATION, OIL PRODUCTION, WATER  
PRODUCTION, FIELD GEOPHYSICAL SURVEYS, TIME SERIES,  
INSULATING REPAIRS, MACHINE LEARNING.**

In the work, various VOI calculation methods were analyzed, and a proprietary VOI calculation method was developed that is suitable for our task. Data was collected and sorted from 30 different production wells of field X. Numerical algorithms for the obtained data were also implemented using machine learning and

statistical analysis methods in the Python programming language. A series of numerical experiments of our predictive models on sorted data from a given field was carried out and the effectiveness of this methodology for calculating VOI, with the possibility of application in other fields, was confi

## СОДЕРЖАНИЕ

Введение.....	6
Глава 1. Обзор существующих методик расчета оценки ценности информации .....	10
1.1. Обзор и анализ существующих решений расчета оценки ценности информации VOI.....	10
1.2. Общая терминология оценки ценности информации VOI.....	16
1.3. Цели и задачи ремонтно-изоляционных работ и оценка их эффективности .....	18
1.4. Резюме и выводы к главе 1 .....	22
Глава 2. Методология расчета оценки экономической целесообразности проведения промысловых геофизических исследований для ремонтно-изоляционных работ .....	23
2.1. Прогнозирование дебитов нефти без использования технологии оценки ценности информации .....	23
2.2. Методология расчета оценки ценности информации VOI в общем виде .....	27
2.3. Математические постановки работы алгоритмов машинного обучения для задачи прогнозирования данных .....	35
2.3.1. Математическая постановка алгоритма градиентного бустинга .....	35
2.3.2. Математическая постановка алгоритма Random Forest.....	41
2.3.3. Математическая постановка алгоритма KNNRegression .....	44
2.3.4 Математическая постановка алгоритма ARIMA .....	48
2.3.5 Математическая постановка алгоритма SARIMA .....	51
2.4. Предварительная подготовка входных данных для прогнозирования расчетных моделей для дебитов нефти и воды .....	54
2.5. Методика расчета показателей ожидаемых ценностей проекта, при наличии дополнительной информации и при отсутствии дополнительной информации .....	58
2.5.1. Алгоритм расчета показателя ценности информации при наличии дополнительной информации .....	58
2.5.2. Алгоритм расчета показателя ценности информации при отсутствии дополнительной информации .....	61

2.6. Резюме и выводы к главе 2 .....	63
Глава 3. Техническая реализация методологии VOI для задачи оценки экономической целесообразности проведения промысловых геофизических исследований для ремонтно-изоляционных на основе исторических данных о скважине .....	64
3.1. Предварительная обработка входных данных для прогнозирования параметров расчетных моделей для дебитов нефти и воды .....	64
3.2. Техническая реализация расчетных алгоритмов машинного обучения для получения параметров ожидаемых ценностей проекта с учетом дополнительной информации .....	71
3.2.1. Техническая реализация расчетных алгоритмов машинного обучения для получения параметра ожидаемой ценности проекта с учетом дополнительной информации .....	72
3.2.2.a. Техническая реализация алгоритма Random Forest, для задачи прогнозирования дебитов нефти и воды .....	72
3.2.3б. Техническая реализация алгоритма KNNRegression, для задачи прогнозирования дебитов нефти и воды .....	75
3.2.3в. Техническая реализация алгоритма градиентного бустинга, для задачи прогнозирования дебитов нефти и воды .....	77
3.2.2. Техническая реализация расчетных алгоритмов машинного обучения для получения параметра ожидаемой ценности проекта при отсутствии дополнительной информации .....	84
3.3 Программная реализация расчетного модуля для задачи оценки экономической целесообразности проведения промысловых геофизических исследования для ремонтно-изоляционных работ на основе исторических данных о скважине .....	87
Заключение .....	96
Список использованной литературы.....	98
Приложение 1 .....	105

## ВВЕДЕНИЕ

Актуальность темы обусловлена тем, что нефтяные ресурсы с каждым годом становится все труднее добывать. Вследствие, чего планировать вложение средств в их добычу становится далеко не тривиальной задачей. Ведь получение максимальной прибыли при минимизации затрат – есть главная задача любой коммерческой организации. В нефтяном инжиниринге есть очень много направлений, где требуется оптимизация производственных затрат. Управление затратами невозможно без уяснения их сущности, знания структуры и специфики их поведения. Цель построения системы управления затратами не сводится к ограничению и контролю расходов, она заключается в оптимизации затрат при получении заданного результата. В данном исследовании мы разработаем методологию оптимизации производственных затрат, на примере определения кандидатов для проведения промысловых геофизических исследований скважин, на основе исторических данных о работе скважин.

В настоящее время прогресс в развитии информационных технологий открывает большие возможности в анализе, разработке и моделировании различных бизнес-процессов на основе различных статистических данных и дает возможность их испытать в «полевых» условиях. В связи с глобальной тенденцией перехода на месторождения со сложным геологическим строением, характеризующим высокую степень неопределенности, подход на основе ценности информации является особенно актуальным. Расчет ценности информации позволяет оценить ценность единичного исследования, найти оптимальную программу и оценить ее достаточность. Кроме того, оценка ценности информации позволяет улучшить понимание рисков проекта и в перспективе составить оптимальную программу разработки. В данной работе метод ценности информации применён для оценки необходимости подбора оптимального планирования проведения ПГИ для РИР. Данный подход призван обеспечить повышение эффективности исследований, снижение трудозатрат на их планирование и повышение прозрачности самих процессов

планирования и обоснованности выбранных для исследования скважин и методов исследования.

В данной работе объектом исследования является нефтегазовая компания, один из лидеров в России по добыче углеводородов. Стратегическая задача предприятия – максимизация добычи углеводородов на месторождениях.

Основные виды работ предприятия – это разведка и разработка нефтегазовых месторождений их добыча и транспортировка, переработка сырой нефти, нефтепродуктов и другой продукции, добыча и реализация нефти, нефтепродуктов и продуктов переработки углеводородов, включая продажу конечным потребителем и экспорт.

Несмотря на то, что финансовое состояние у фирмы хорошее, компания стремится к уменьшению издержек и увеличению прибыли. Но на месторождениях, где тяжело добывать нефть, необходима очень сложная и дорогостоящая инфраструктура по транспортировке и эксплуатации скважин, что очень сильно повышает расходы и уменьшает финансовую отдачу. Отсутствие хорошо продуманной методологии и технологии по оценке экономической целесообразности проведения промысловых геофизических исследований и предопределило содержание научной работы.

Объект исследования – заключения промысловых геофизических исследований с 30 добывающих скважин и наборы данных о работе этих скважин с месторождения X.

Предмет исследования – оптимизации финансовых издержек, при грамотном определении скважин-кандидатов на проведение промысловых геофизических исследований.

Цель исследования – на основе теоретического и практического опыта разработать методологию оценки экономической целесообразности проведения промысловых геофизических исследований на основе исторических данных о работе скважин и заключений ПГИ месторождения X.

Задачи исследования:

– Анализ существующих методик по оптимальному планированию, которые



базируются на основе исторических данных.

– Разработка методологии и технологии по оценке экономической целесообразности проведения ПГИ на основе исторических данных о скважине.

– Оценить эффективность разработанной методологии и технологии.

– Разработать программный расчетный комплекс, базирующийся на проведенном исследовании.

Теоретической основой исследования явились теории и концепции, сформированные в трудах отечественных и зарубежных исследователей, которые занимались оценкой экономической эффективности, на основе исторических данных.

Методологической основой исследования послужил системный подход к изучению оптимизационных процессов, осуществляемых в сфере нефтяного инжиниринга. В процессе исследования применялись методы общенаучного познания, научной абстракции, анализа и синтеза, сопоставления и сравнения. Методологию исследования составили как общенаучные методы (прогнозирование, моделирование), так и специальные методы (графический и статистический анализы, методы машинного обучения и др.).

Информационную базу исследования составили заключения промысловых геофизических исследований, исторические данные о работе добывающих скважин месторождения X, монографии по теории оптимального экономического планирования, монографии по статистическому анализу, зарубежные исследование по тематике машинного обучения, материалы периодических изданий и ресурсы сети Интернет.

Научная новизна исследования состоит в разработке теоретических основ оценки экономической целесообразности на основе исторических данных о скважине, используя технологии машинного обучения и статистического анализа данных. А также реализация программного комплекса для автоматизации производственного процесса, что позволяет уменьшать издержки производства.



## **ГЛАВА 1. ОБЗОР СУЩЕСТВУЮЩИХ МЕТОДИК РАСЧЕТА ОЦЕНКИ ЦЕННОСТИ ИНФОРМАЦИИ.**

### **1.1 Обзор и анализ существующих решений расчета VOI.**

Одним из основных элементов контроля разработки нефтяных залежей являются промыслово-геофизические исследования (ПГИ) скважин.

Результаты проведения промыслово-геофизических исследований обеспечивают технологический, эксплуатационный, геолого-промысловый и технический контроль разработки месторождений. Задача технологического контроля включает выбор оптимального режима работы скважины и её технологического оборудования. Геолого-промысловый контроль включает исследования процессов вытеснения нефти и газа в пласте. Для оценки технического состояния скважин контролируют качество и целостность цементного камня, проводится оценка герметичности колонн, глубины текущего искусственного забоя и др. Результаты интерпретации ПГИ позволяют обосновывать и контролировать геолого-технологические мероприятия (ГТМ), позволяющие регулировать разработку месторождений и поддерживать плановые показатели добычи.

Следовательно, достоверные результаты промысловых геофизических исследований скважин являются важнейшим источником информации о характеристиках состояний скважин. В связи с этим выбор и реализация ГТМ должны основываться на основе данных ПГИ, но полный охват действующего фонда скважин промыслово-геофизическими исследованиями в большинстве случаев затруднён из-за технологических ограничений и затрат на проведение исследований, в том числе за счёт потерь добычи нефти при остановке скважин на время проведения ПГИ.

Существует обширный список существующих технологий ПГИ, выбор которых зависит от различных факторов: взаиморасположение насосно-компрессорных труб (НКТ) и интервалов перфорации, число совместно эксплуатируемых пластов, дебит/расход закачиваемой жидкости, газосодержание продукции и др. В зависимости от технологий и методов проведения ПГИ будут зависеть затраты на их проведение, в том числе за счёт потерь добычи нефти при остановке скважин на время проведения ПГИ.

Принятие решения о выборе скважины для проведения ПГИ, предполагает и экономическое обоснование, что является нетривиальной задачей в условиях неопределенности. При планировании ПГИ необходимо определять баланс между затратами на проведение исследования и ценностью информации, которую даёт исследование.

В рамках НИР был разработан подход к оптимальному планированию проведения ПГИ на основе методики экономической оценки ценности информации (VOI) на основе исторических данных со скважин.

Проанализируем исследования по анализу ценности информации (VOI) и его практическое применение к различным приложениям, основываясь на обзоре публикаций SPE, практическое рассмотрение будет приведено на примере нефтегазовой отрасли. Будут рассмотрены возможные причины ограниченного использования методов VOI и предлагаемые способы расширения использования этого инструмента в анализе ценности информации и сформирована новая практическая задача с приложением к нефтяной промышленности.

Впервые концепция VOI была представлена в нефтегазовых исследованиях в 1960, ученым К. Гейсоном [23]. Эта концепция хорошо известна среди инженеров-нефтяников и геологов. Фактически, инженеры и ученые склонны полагать, что большое количество информации вносит вклад в принятие стратегических важных решениях, так как информация уменьшает неопределенность при принятии решений. Действия нефтяных инженеров или

геологов, заключаются в том, чтоб собрать большее количество информации, с целью улучшения процесса принятия решений. Основной причиной сбора информации или проведение любого технического анализа является устранение неопределенностей для улучшения оценки принятых решений. Любого процесс сбора информации заключается в том, стоит ли улучшение процесса принятия решений получение информации затрат на их принятие. Методология VOI помогает разобраться в данной проблеме.

Был проведен анализ исследований методологии VOI с приложениями к нефтегазовой промышленности. Искались термины «Ценность информации», «Ценность данных», «значение сеймики», «значение 3D сеймики», «Значение 4D сеймики», «значение каротажа», «значение активной зоны» и «Ценность скважины». Было найдено 30 статей VOI, опубликованных в литературе SPE.

Шлайфер в 1959 определил концепцию VOI в контексте анализа ценности информации деловых решений [58]. Его книга сосредоточена на проблемах статистического вывода и выборки, но не предоставляет первое теоретическое определение VOI, но обсуждает проблему принятий решений в условиях неопределенности информации в широком смысле слов. Важность публикации Шлайфера не может быть переоценена. Как заявил Хиршлайфер в 1961 «Работа Шлайфера, на мой взгляд, не слишком рекомендуется студентам или практикующим специалистам; что он сделал, это структурирование в набор операционных процедур группу революционных идей, которые, подрывая старые порядки статистического вывода, но так и не дал практикующим или потребителям статистики что-нибудь, чтобы заменить старый порядок " [31, стр. 471-489].

Теория современной субъективной вероятности показывает то, что, когда знания или информация бесплатны, они платят в ожидании приобрести его [53]. Однако эта теория не могла количественно и качественно оценить

ценность информации, так как ей не хватало оценки последствий и не включала в себя принятие решений. Только через год после того, как Шлайфер определил концепцию VOI в бизнес-аналитическом контексте, Гейсон представил его для нефтегазовой отрасли в своей книге «Решения в условиях неопределенности: Решения по бурению нефтяных и газовых операторов» [23]. В этом исследовании Гейсон представил организацию процедур по принятию стратегических решений. На примере буровых решений в главе «Информационные решения» подробно иллюстрирует, как использовать концепцию VOI, чтобы помочь инженерам при «покупке информации». Книга Гейсона является перепечаткой его диссертации из Гарвардской школы бизнеса. Но термин «ценность информации» был представлен только спустя 23 года после исследования Гейсона в 1983 в статье Уоррена [68]. Обсуждаемые выше публикации легли в основу последующих документов SPE по анализу оценки ценности информации. В анализируемых исследованиях часто повторяются ранее опубликованные идеи и отсутствует четкое руководство о том, как на самом деле стоит использовать концепции VOI.

Обсудим одну теоретическую статью. В статье Мораса и его коллег использовалась концепция VOI для разработки оптимальной системы мониторинга газохранилищ [43]. Их методология была достаточно продвинутой, но в итоге исследователи полагались на эвристику при размещении контрольных скважин. Последние достижения в вычислительных способностях современных компьютерных систем и алгоритмов помогают в решении этих проблем.

**Надежность:** критическим вкладом в анализ VOI является оценка достоверности информации. Пять из исследуемых работ иллюстрируют идеальную информацию, которая не нуждается в дополнительной оценке. Остальные документы касаются с неинформативных данных. Только 13 из 25 работ посвящены достоверности информации. Из них 11 оценивают

функцию правдоподобия или вероятности через экспертную оценку. Использование экспертной оценки рассматривается как недостаток концепции VOI. В двух статьях попытка выйти за рамки экспертных оценок и связать оценку ценности информации с функцией правдоподобия для некоторых исследовательских моделей (например, геофизическая) [7],[5].

**Байесовская модель:** любой анализ VOI требует байесовского расчета. В простейшей ситуации строится дискретное дерево вероятностей, оценивает вероятности и выполняет байесовский расчет. Эта техника используется в почти в каждой статье, где имеет место данные плохого качества. Этот метод плохо работает, если есть необходимость изучать модели непрерывных свойств коллекторов (например: пористость). Для обработки постоянных неопределенностей используется сопряженное предварительное распределение, чтобы соответствовать процессу отбора проб. Для ситуаций, когда нет сопряженных распределений приоритетней использовать метод цепей Маркова или метод Монте-Карло.

**Источник информации:** в большинстве исследований по VOI ориентированы на оценку сейсмической информации. Только одно исследование [69] дает количественную оценку стоимости взятие образца ядра. В большинстве работ рассматривается только один источник информации для оценки проведения анализа VOI. В двух исследованиях [18],[69] рассматривается возможность выбора информации из различными источниками или последовательностью информации из различных источников.

**Моделирование решений:** информация должна содержать возможность изменение принятия решений, если оно имеет ценность. Это требует явного моделирования оценки ценности решений, на которые влияют неинформативные данные. Большинство рассмотренных исследований, не делают это. Вместо этого в исследованиях предполагается, улучшение в оценке информации произойдет после сбора информации. В большинстве случаев информация меняет наши убеждения в оценке ценности информации, поэтому убеждения должны быть отделены от оценки принятия

решений.

**Оценка достоверности информации:** при наличии несколько источников информации анализ принятия решения должен включать оценки коррелированных распределений функций правдоподобия. Эти параметры тяжело определять поэтому инженеры пытаются оценить при помощи эвристики. Самый сложный и необычный параметр для оценки может быть функция вероятности, так как отсутствует традиция сбора информации в этом контексте. Очень сложно измерить точность или надежность оценочных скважин, сложно оценить исследование по моделированию коллекторов или каротажных интерпретаций. Это количественное определение, требуется для анализа VOI. Следовательно, исследователям необходимо научиться это делать. Если имеется доступ к соответствующим историческим (или другим) данным, то они должны быть использованы при оценке информации. В оценке качества информации из сейсмических исследований или анализа ядра, можно выделить соответствующие исторические данные. Большинство нефтегазовых компаний имеют долгую историю сбора этого типа данных, но очень немногие разработали подходящие базы данных достоверности информации, которые они собирают. Такие базы данных не только облегчат запуск применение методологии оценки информации, но также будут способствовать обеспечению согласованности между несколькими анализами ценности информации. Требуется много исследований для разработки подходящих моделей для мероприятий по сбору и оценке информации.

**Реализация и программные приложения:** многие исследования базируются на простых моделях. Это хорошо работает для относительно простых ситуаций, когда аналитик имеет большой опыт в разработке моделей VOI. Широкое и последовательное использование концепции VOI потребует улучшения в развитии методологии VOI, а также улучшенных программных приложений, нацеленных на анализ принятия решений. Таким приложениями следует сосредоточиться на простом структурировании решений проблем



оценки информации и на алгоритмах принятия решений. Приложения должны включать более сложные байесовские модели, или могут быть включены такие методы, как моделирование марковских цепей, метод Монте-Карло и современные технологии машинного обучения и анализа данных. Приложение, поддерживающее принятие решений, должно включать в себя возможность исследования надежности выбора решений оценки VOI в зависимости от изменений на входе параметров, включающие предшествующие вероятности.

## Поменять порядок

### 1.2 Общая терминология оценки ценности информации VOI.

Метод VOI (Value of information) в нефтегазовой промышленности впервые введён в 1960 году Джексоном Гейсоном [23]. Ценность информации (VOI) – это разница между ожидаемой ценностью проекта с учетом и без учета полученной информации.

$$VOI = EMV_{wi} - EMV_{woi} \quad (1.1)$$

$$EMV_{wi} = \sum P_i \cdot NPV_{wi} \quad (1.2)$$

$$EMV_{woi} = \sum P_i \cdot NPV_{woi} \quad (1.3)$$

Где  $EMV_{wi}$  – ожидаемая ценность проекта с учетом дополнительной информации;

$EMV_{woi}$  – ожидаемая ценность проекта без учета дополнительной информации;

$NPV_{wi}$  – сумма дисконтированных значений потока платежей, приведенных к сегодняшнему дню с учетом дополнительной информации;

$NPV_{woi}$  – сумма дисконтированных значений потока платежей, приведенных к сегодняшнему дню, без учета дополнительной информации;

$P_i$  – вероятность того, что событие произойдет.

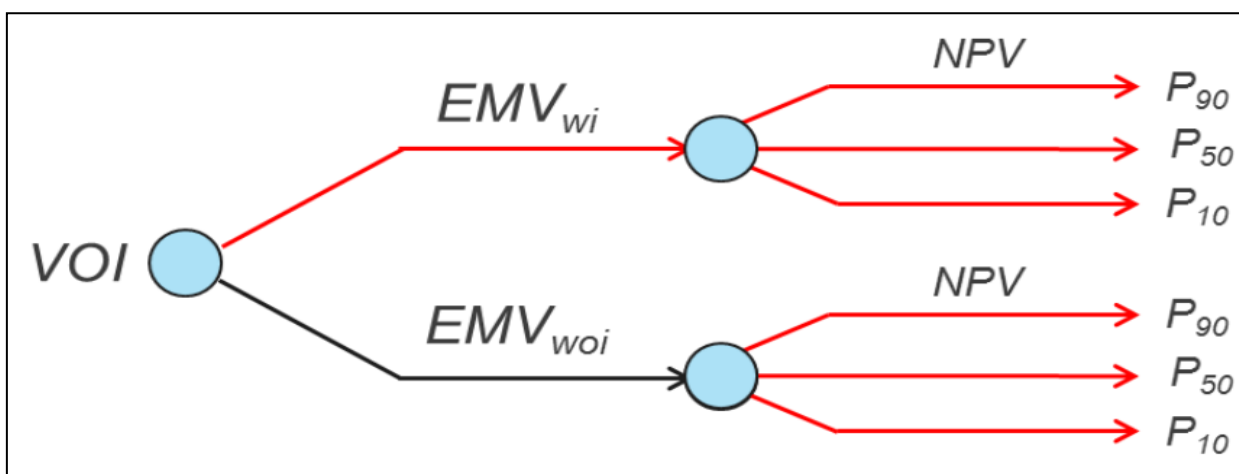


Рис. 1.1 Общий вид дерева решений для расчета оценки ценности информации VOI.

Оценка экономической ценности информации позволяет охарактеризовать экономическую выгоду от выполнения исследования, произвести оценку эффективности программы исследовательских работ, а также оценить достаточность исследований на основе дерева решений (Рис.1.1). Дерево решений – это графический метод анализа информации, который отображает процесс принятия управленческих решений в виде набора последовательных альтернативных действий и реализацией сценариев характеризующиеся наличием рисков и неопределенностей, комбинации которых, при заданных вероятностях, приводят к достижению поставленных целей. Из-за глобальной тенденцией перехода на месторождения со сложным геологическим строением, характеризующим высокую степень неопределенности информации, подход на основе оценки ценности информации является особенно актуальным. Расчет VOI позволяет оценить ценность единичного исследования, а также найти оптимальную программу исследований и оценить достаточность проведения исследований. Оценка ценности информации позволяет улучшить понимание рисков проекта и помочь в составлении оптимальной программы разработки месторождений.

В данной исследовании методологию VOI решено было исследовать для оценки необходимости подбора оптимального планирования проведения промысловых геофизических исследований для ремонтно-изоляционных

работ. В данном исследовании будет обеспечено повышение эффективности исследований, снижение трудозатрат на их планирование и повышение прозрачности самих процессов планирования и обоснованность выбора данных для исследований скважин. Проведем анализ переменных, которые могут оказать влияние на качество информативности исследуемых данных при экономической оценке информации подбора оптимального проведения промысловых геофизических исследований для ремонтно-изоляционных работ.

### **1.3 Цели и задачи ремонтно-изоляционных работ и оценка их эффективности.**

Произведем оценку влияния ремонтно-изоляционных работ на информативность данных. Ремонтно-изоляционные работы (РИР) – мероприятия по преграждению путей проникновения вод в скважину, а также отключение от неё отдельных пластов и обводнённых интервалов. Данные работы – одно из главных средств увеличения объёмов извлеченной нефти и газа из пласта, что оказывает влияние на качество информации набора данных.

Ремонтно-изоляционные работы проводят с целью изоляции верхних вод, нижних вод, поступающих через цементный стакан и по заколонному пространству, подошвенных и контурных вод, поступающих по наиболее проницаемым интервалам и трещинам пласта, т.е. обеспечивают оптимальные условия работы продуктивного пласта, для достижения запланированной (максимальной) выборки запасов нефти

Определение необходимых объемов проведения РИР в фонде нефтяных скважин является лишь частью задачи планирования этих работ. Планирования РИР и планирование проведения любого вида ГТМ, не имеет смысла без наличия данных об эффективности планируемых работ. Техно-экономические показатели являются определяющими при составлении плана

проведения указанных работ. При этом эффект от проведения РИР является одним из основных показателей самого плана их проведения.

Методики оценки эффективности РИР в данное время не существует. Это относится к оценке технологического эффекта от проведения РИР, выражаемого в определенных количественных показателях. Ниже рассмотрены основные методические положения по оценке технологического эффекта от проведения РИР в добывающих скважинах, а также рекомендуемая методика оценки технологического эффекта в отдельных нефтяных скважинах.

Оценку качества успешности этих работ проводят с помощью методов гидродинамических и промыслово-геофизических исследований.

Успешность РИР по отключению отдельных интервалов пластов оценивается по данным исследования профиля притока в нефтяных скважинах глубинными дебитомерами или профиля приемистости в нагнетательных скважинах глубинными расходомерами. Степень снижения продуктивности самого пласта может быть оценена и по кривым восстановления давления.

Успешность РИР по исправлению некачественного кольца оценивается по результатам специальных исследований с помощью методов: АКЦ, термометрии, радиоактивных изотопов и т. д.

При проведении большинства РИР перечисленные выше исследования проводят по ограниченному фонду ремонтируемых скважин. В этих случаях оценивают качество проведенных РИР по полученному технологическому эффекту, что не всегда равнозначно вследствие несоответствия представлений действительному характеру обводнения скважины. Оценка технологической эффективности проводимых РИР сопряжена с рядом трудностей, определяемых исключительным многообразием и сложностью явлений,

взаимообусловливаемых и являющихся следствием как самих РИР, так и вызванных причинами, не связанными с ними.

При сохранении режима работы скважин, успешно выполненные РИР приводят к уменьшению дебита жидкости нефтяных и приемистости нагнетательных скважин, что обусловлено уменьшением продуктивности скважин вызванным сокращением вскрытой толщины продуктивного пласта или ликвидацией нарушений в цементном кольце или колонне. Именно указанные причины обуславливают необходимость после проведения РИР в ряде случаев изменять режим работы скважины в сторону уменьшения добычи нефти.

Достижение поставленной перед РИР цели во всех случаях приводит к изменению фильтрационных потоков движения жидкости в продуктивных пластах, обеспечивающему в добывающих скважинах снижение притока воды и сохранение или увеличение притока нефти.

При оценке экономической целесообразности того или иного вида РИР должен учитываться суммарный технологический эффект от их проведения, т. е. за весь период его проявления.

Ниже приведен порядок определения технологического эффекта осуществления РИР в отдельных скважинах по изложенной методике и для сравнения по существующим методикам: с учетом и без учета изменения величин дебита нефти и содержания воды в ремонтируемых скважинах.

При этом использовали следующие зависимости оценки эффективности РИР:

Методика 1 предусматривающая сохранение величин дебита нефти и воды без проведения РИР на уровне последних трех месяцев перед РИР.

Методика 2, предусматривающая изменение величин дебита нефти и воды без проведения РИР в соответствии с коэффициентами их изменения перед РИР.

Методика 3 предусматривающая изменение величины дебита нефти" без проведения РИР в соответствии с коэффициентом его изменения перед РИР и сохранение величины обводненности добываемой нефти на уровне последних трех месяцев перед РИР (рекомендуемая методика).

В указанных условиях для оценки эффективности проводимых РИР рекомендуется использование методики под номером 3, как наиболее полно учитывающей изменение режима эксплуатации скважины до и после ремонта.

#### **1.4 Выводы к главе 1.**

Концепция методологии VOI – является отличным инструментом для лиц, чья сфера деятельности связана с применением технологий оценки принятия решений. Она имеет хорошее математическое представление и является простой по собственной структуре. VOI может помочь определить потенциальную ценность любой информации, а также ценность улучшенного приобретения технологии, до фактического сбора информации. Обзор литературы SPE методологий VOI показывает, что нефтяная индустрия не сделала VOI неотъемлемой частью процесса принятия решений, но используется при определенных бизнес-решениях. Но для задачи экономической оценки целесообразности проведения промысловых геофизических исследований на основе исторических данных о скважине нет методологии расчета VOI. Поэтому было решено разработать методологию расчета оценки ценности информации для этой задачи и реализовать её алгоритмы в расчетном модуле.

## **ГЛАВА 2. МЕТОДОЛОГИЯ РАСЧЕТА ОЦЕНКИ ЭКОНОМИЧЕСКОЙ ЦЕЛЕСООБРАЗНОСТИ ПРОВЕДЕНИЙ ПРОМЫСЛОВЫХ ГЕОФИЗИЧЕСКИХ ИССЛЕДОВАНИЙ ДЛЯ РЕМОНТНО-ИЗОЛЯЦИОННЫХ РАБОТ.**

### **2.1 Прогнозирование дебитов нефти без использования технологии оценки ценности информации.**

Предсказание добычи нефти по скважинам – ответственная задача, так как это основа для роста нефтяной промышленности в целом.

В планировании добычи нефти используют 2 метода прогнозирования: гидродинамические методы и статистические методы прогнозирования.

Гидродинамические методы прогнозирования используют при анализе разработки продуктивных нефтяных пластов. При использовании этого метода делается предсказание на основе фактов о геологическом и физическом свойстве объектов.

Методы гидродинамического прогнозирования зависят от большого числа факторов, которые оказывают влияние на прогноз дебита нефти. Влияние этих факторов может сказываться негативно на качестве прогноза дебита нефти.

Статистические методы прогнозирования основываются на статистическом анализе данных за прошедший промежуток времени с дальнейшим экстраполированием на будущий промежуток времени.

Использование статистических методов прогноза дебитов нефти достаточно ограничен. Чтоб определить различные зависимости и корреляции, необходим очень продолжительный промежуток по эксплуатации скважин, что не всегда возможно на практике.

На данный момент при прогнозировании добычи нефти по месторождениям базовым методом является гидродинамическое моделирование, а корректировки и уточнения осуществляются статистическими приемами прогнозирования исходя из исторических данных и текущих состояний

скважин.

При прогнозировании за базовые входные данные берутся данные из технологических схем и данных разработки нефтяных залежей. Дебит нефти на промежуток времени, в который планируется делать предсказание считается как суммарный дебит нефти «старых» и новых скважин:

$$Q_{t+1} = Q_{c(t+1)} + Q_{h(t+1)} \quad (2.1)$$

где  $Q_{t+1}$  – добыча нефти в планируемом году;

$Q_{c(t+1)}$  – добыча нефти в планируемом году из перешедших с прошлого года (старых) скважин;

$Q_{n(t+1)}$  – добыча нефти в планируемом году из новых скважин, которые предусмотрено ввести в эксплуатацию из бурения, а также из освоения с прошлых лет;

$t+1$  — индекс планируемого года;

$t$  — индекс года, предшествующего планируемому.

Первая составляющая формулы (4) - добыча нефти из переходящих (старых) скважин - определяется из соотношения

$$Q_{c(t+1)} = (Q_{ct} + q_t \cdot N_t \cdot U_{t+2}) \cdot K_{u(t+1)} \quad (2.2)$$

$Q_{c(t+1)}$  – добыча из старых скважин в предыдущем году;

$q_t, N_t, U_{t+2}$  – средний дебит, количество новых скважин, введенных в эксплуатацию в предшествующем году, и число дней работы каждой из них в планируемом году;

$K_{u(t+1)}$  - коэффициент изменения добычи нефти в планируемом году.

Величина  $(Q_{ct} + q_t \cdot N_t \cdot U_{t+2})$  представляет собой расчетную добычу, которая имела бы место в планируемом году, если бы добыча из скважин прошлого года осталась без изменения, а новые скважины, введенные в прошлом году, эксплуатировались бы в планируемом году при тех же дебитах полное число дней.

Как правило, определить эту расчетную добычу, имея фактические данные



(добычу из старых скважин, количество и средний дебит новых скважин) за предшествующий год, не представляет трудностей.

При этом число дней работы переходящих новых скважин  $U_{t+1}$  принимается равным:

$$U_{t+1} = 365 \cdot K_9 \quad (2.3)$$

$K_9$  – коэффициент эксплуатации.

Расчет коэффициента добычи нефти  $K_{u(t+1)}$  является самым сложным расчетным параметром, так как неизвестно какое количество скважин будет введено в эксплуатацию в предстоящем году.

При вытеснении нефти из залежей водой происходит падение уровня нефти при добыче, так как при увеличении уровня воды доля нефти уменьшается. Также идет падение уровня дебита нефти из скважин и уменьшение добычи из «старых» скважин месторождения.

Падение в перспективе дебита по жидкости связано с падением или увеличением давления в нефтяных пластах, с изменением продуктивности в скважинах из-за различного рода засорений или очищения призабойной зоны в скважинах, и от неправильного разбуривания скважин на месторождениях. Следственно после проведения различного рода мероприятий средний дебит по жидкости несильно растет, а к концу эксплуатации скважины снова начинает падать, так как происходит выбывание из фонда высокодебитных обводненных скважин.

Изменение фонда старых скважин бывает связано главным образом с выходом скважин из эксплуатации из-за обводнения в связи с подходом вытесняющей нефть воды, переводом части скважин под нагнетание воды и др.

Каждый из параметров в отличие от коэффициента изменения добычи нефти может быть определен гидродинамическими методами прогнозирования или методами статистического анализа, исходя из стадии и текущего состояния разработки залежи, а также на основании обобщения опыта разработки по аналогии с другими подобными месторождениями.

Вторая составляющая формулы (2.1) - добыча нефти из новых скважин, которые будут введены в эксплуатацию в планируемом году, рассчитывается по произведению:

$$Q_{n(t+1)} = N_{n(t+1)} \cdot q_{n(t+1)} \cdot m_{(t+1)} \quad (2.4)$$

$N_{n(t+1)}$  — число вводимых в эксплуатацию в планируемом году новых добывающих скважин;

$q_{n(t+1)}$  — проектируемый среднесуточный дебит нефти новых скважин в планируемом году;

$m_{(t+1)}$  — среднее число суток работы одной новой добывающей скважины в планируемом году.

При расчете добычи из новых скважин наиболее сложно прогнозировать их средний дебит, так как в эксплуатацию последовательно разбуриваются участки с различной продуктивностью

По эксплуатируемым месторождениям размещение проектных точек производится на карте разработки, и для каждой из них дебит устанавливается исходя из дебитов окружающих действующих скважин.

По новым месторождениям или при вводе в разработку неразбуренных участков эксплуатируемых месторождений дебиты новых скважин целесообразно определять исходя из их вероятной продуктивности по картам нефтенасыщенной толщины или гидропроводности.

Среднее число дней работы одной новой добывающей скважины в планируемом году рассчитывается исходя из месячного распределения ввода новых скважин в эксплуатацию, выполняемого с учетом планируемых сроков окончания строительства скважин, нормативной продолжительности их обустройства и освоения, а также сезонных условий проведения этих работ. Обычно число дней работы новых скважин планируемого года составляет от 130 до 150.

Описанная методология расчета прогнозов дебитов нефти требует очень много неизвестных переменных. Для того чтобы спрогнозировать дебит нефти

необходимо произвести множество трудоемких расчетов, геофизических и гидродинамических исследований и симуляций для получения необходимых параметров расчета дебита. Производство этих работ является дорогостоящим и трудоемким процессом, а суммарный финансовый эффект отдачи от исследований может оказаться не рентабельным. Поэтому появилась необходимость в новых методиках прогнозирования нефти, которая является менее дорогостоящей и менее трудоёмкой, но эффективной. В данном исследовании будет разработана данная методология на примере задачи экономической целесообразности проведения промысловых геофизических исследований скважин для ремонтно-изоляционных работ на основе исторических данных о скважине.

## **2.2 Методология расчета оценки ценности информации VOI в общем виде.**

Для задачи оценки экономической целесообразности промысловых геофизических исследований для ремонтно-изоляционных работ отсутствует методы оценки информации. Вследствие чего было решено разработать методологию расчета VOI для этой задачи. В соответствии с тем что было описано в главе 1, будем продвигаться от более высокого уровня абстракции к низкому уровню абстракции. Рассмотрим основные компоненты расчета VOI. Основная формула расчета выглядит следующим образом.

$$VOI = EMV_{wi} - EMV_{woi} \quad (2.5)$$

Где  $EMV_{wi}$  – ожидаемая ценность проекта с учетом дополнительной информации;

$EMV_{woi}$  – ожидаемая ценность проекта без учета дополнительной информации;

Это первый уровень абстракции. Спустимся на уровень абстракции ниже рассмотрим каждый из параметров.

$$EMV_{wi} = \sum P_i \cdot NPV_{wi} \quad (2.6)$$

$$EMV_{woi} = \sum P_i \cdot NPV_{woi} \quad (2.7)$$

$NPV_{wi}$  – сумма дисконтированных значений потока платежей, приведенных к сегодняшнему дню с учетом дополнительной информации;

$NPV_{woi}$  – сумма дисконтированных значений потока платежей, приведенных к сегодняшнему дню, без учета дополнительной информации;

$P_i$  – вероятность того, что  $i$ -й сценарий произойдет.

В формулах (2.6) и (2.7) содержится знак суммы, это означает сумму всех возможных сумм дисконтированных значений потока платежей к сегодняшнему дню с учетом и без учета дополнительной информации. Вероятности для каждого месторождения высчитываются статистически, но также их значения могут задаваться экспертами.

Спустимся на уровень абстракции ниже и проанализируем все возможные развития сценариев  $NPV$  для ожидаемых ценностей проекта с учетом дополнительной информации и без её учета. Рассмотрим сценарии, которые включают учет дополнительной информации. Разработанная модель предусматривает 6 возможных варианта развития событий ( $NPV1 - NPV6$ ).

$NPV1$  предполагает исход, при котором был проведен стандартный комплекс ПГИ, который выявил отсутствие необходимости, либо невозможности проведения РИР.

$NPV2$  предполагает исход, в результате которого был проведен стандартный комплекс ПГИ, оказавшийся недостаточно информативным вследствие чего было проведено последующее расширенное ПГИ, которое выявил отсутствие необходимости, либо невозможность проведения РИР.

$NPV3$  предполагает исход, при котором был проведен стандартный комплекс ПГИ, при котором появилась необходимость проведения РИР.

*NPV4* предполагает исход, при котором был проведен стандартный комплекс ПГИ. Первичный ПГИ оказался не информативным, вследствие чего был проведен расширенный комплекс ПГИ, в результате которого появилась необходимость проведения РИР.

*NPV5* предполагает исход, при котором был проведен стандартный комплекс ПГИ, в результате которого появилась необходимость в проведения РИР, затем проведен еще один стандартный комплекс ПГИ, вследствие чего появилась необходимость проведения повторного комплекса РИР.

*NPV6* предполагает исход, при котором был проведен стандартный комплекс ПГИ, в результате которого появилась необходимость в проведения РИР, затем проведен расширенный комплекс ПГИ, при котором также появилась необходимость проведения комплекса повторных РИР.

Ожидаемая ценность проекта без информации (*NPV7*), в случае не проведения ПГИ, сводится к прогнозу дебита нефти. Математическое представление сценариев *NPV* :

$$NPV1 = Q_n(t)'C_n - Q_g(t)'C_g - C_{nzu} - Q_n(t_{nzu})'C_n \quad (2.8)$$

$$NPV2 = Q_n(t)'C_n - Q_g(t)'C_g - C_{nzu} - Q_n(t_{nzu})'C_n - C_{nzu2} - Q_n(t_{nzu2})'C_n \quad (2.9)$$

$$NPV3 = Q_n(t)''C_n - Q_g(t)''C_g - C_{nzu} - Q_n(t_{nzu})'C_n - C_{rip} - Q_n(t_{rip})C_n - C_{nzu}' - Q_n(t_{nzu}')C_n \quad (2.10)$$

$$NPV4 = Q_n(t)''C_n - Q_g(t)'C_g - C_{nzu} - Q_n(t_{nzu})'C_n - C_{nzu2} - Q_n(t_{nzu2})'C_n - C_{rip} - Q_n(t_{rip})C_n - C_{nzu}' - Q_n(t_{nzu}')C_n \quad (2.11)$$

$$NPV5 = Q_n(t)''C_n - Q_g(t)''C_g - C_{nzu} - Q_n(t_{nzu})'C_n - C_{rip} - Q_n(t_{rip})'C_n - C_{nzu}' - Q_n(t_{rip})'C_n - C_{nzu}' - Q_n(t_{nzu}')'C_n - C_{rip2} - Q_n(t_{rip2})'C_n - C_{nzu}'' - Q_n(t_{nzu}'')'C_n \quad (2.12)$$

$$NPV6 = Q_n(t)''C_n - Q_g(t)''C_g - C_{nzu} - Q_n(t_{nzu})'C_n - C_{nzu2} - Q_n(t_{nzu2})'C_n - C_{rip} - Q_n(t_{rip})'C_n - C_{nzu}' - Q_n(t_{nzu}')'C_n - C_{nzu}' - Q_n(t_{nzu}')'C_n - C_{rip2} - Q_n(t_{rip2})'C_n - C_{nzu}'' - Q_n(t_{nzu}'')'C_n \quad (2.13)$$

$C_n$  – предполагаемый доход от 1 тонны нефти. Денежная величина, которое предприятие заработает, при продаже 1 тонны нефти. Эта величина известна

и постоянна, изменение величины параметра зависит от изменения ситуации на рынке купли-продажи нефтяной продукции.

$C_e$  – предполагаемый расход на транспортировку 1 тонны воды. Денежная величина, которое предприятие израсходует на транспортировку 1 тонны воды, Величина параметра определяется подрядчиками с учетом рыночной стоимости затрат на транспортировку 1 тонны воды.

$C_{ngu}$  – предполагаемый расход на проведение ПГИ. Денежная величина, которую подрядчик израсходует на первичное промысловое геофизическое исследование.

$C_{ngu2}$  – предполагаемый расход на проведение повторного ПГИ. Денежная величина, которая определяется подрядчиком на проведения повторного промыслового геофизического исследования, если первичное ПГИ оказалось неинформативным или требуются дополнительное получение данных разновременных измерений для их последующего сравнительного анализа и выявления на этой основе характеристик флюидов в стволе скважины, в продуктивных пластах и эксплуатируемых залежей.

$C_{rip}$  – предполагаемые расходы на проведение РИР. Денежная величина, которую подрядчик израсходует на проведение ремонтно-изоляционных работ

$C_{rip2}$  – предполагаемые расходы на проведения повторного РИР. Денежная величина, которую подрядчик израсходует на проведения повторных ремонтно-изоляционных работ, в случае неэффективности первичных ремонтно-изоляционных работ.

$C_{ngu'}$  – предполагаемый расход на проведения ПГИ после первичного РИР. Денежная величина, которую подрядчик израсходует на комплекс промысловых геофизических исследований, после первичных ремонтно-

изоляционных работ. ПГИ проводится для определения качества состояния скважины и обсадных колон после ремонтно-изоляционных работ.

$C_{ngu}$  – предполагаемый расход на проведение ПГИ после вторичных РИР. Денежная величина, которую израсходует подрядчик при проведении промысловых геофизических исследований после вторичных ремонтно-изоляционных работ. Они производятся в случае, если первичные ремонтно-изоляционные работы оказались неэффективными.

$t_{ngu}$  – предполагаемая продолжительность проведения первичных ПГИ. Время, которое понадобилось подрядчику на проведение промысловых геофизических исследований. Величина времени измеряется в сутках и ее изменение зависит от условий района месторождения и видов геологических исследований скважины.

$t_{ngu2}$  – предполагаемая продолжительность проведения повторного ПГИ. Время, которое подрядчик затратит на проведение повторных промысловых геофизических исследований, если первичные исследования окажутся неинформативными. Величина времени измеряется в сутках и ее изменение зависит от условий района месторождения и видов геологических исследований скважин.

$t_{rip}$  – предполагаемая продолжительность проведения РИР. Время, которое затратит подрядчик на проведение ремонтно-изоляционных работ скважины. Величина времени измеряется в сутках и ее изменение зависит от условий района месторождения и видов геологических исследований скважин.

$t_{rip2}$  – предполагаемая продолжительность проведения повторного РИР. Время, которое понадобится подрядчику на проведение вторичных ремонтно-изоляционных работ, случае если первичный комплекс ремонтно-изоляционных работ скважины оказался неэффективным.

$t$  – продолжительность добычи дебитов нефти и воды в месяцах по которому планируется определять экономическую оценку ценности информации (VOI). По умолчанию величина времени, по которому планируется определять дебиты равна 6 месяцам.

$P_i$  – вероятность происхождения  $i$ -го сценария. Выбор сценария определяется по историческим данным скважины, исходя из комплексов проведенных промысловых геофизических исследований и комплексов, проведенных ремонтно-изоляционных работ. Величина определяется экспертной оценкой при помощи методов статистического анализа.

$Q_n$  – посуточная история дебита нефти.

$Q_{ж}$  – посуточная история дебита жидкости.

$Q_n(t)'$  – прогноз накопленного за время  $t$  дебита нефти при текущем состоянии скважины. Прогноз дебита нефти осуществляется на использовании данных дебитов нефти со скважины до проведения различных комплексов промысловых геофизических исследований и комплексов ремонтно-изоляционных работ.

$Q_w(t)'$  – прогноз накопленной за время  $t$  дебита воды при текущем состоянии скважины. Прогноз дебита воды осуществляется на использовании данных дебитов воды со скважины до проведения различных комплексов промысловых геофизических исследований и комплексов ремонтно-изоляционных работ.

$Q_n(t)''$  – прогноз накопленного за время  $t$  дебита нефти после проведения РИР. Прогноз дебита нефти осуществлялся на использовании данных дебитов нефти после проведения ремонтно-изоляционных работ до момента сбора данных либо до проведения вторичных ПГИ.



$Q_w(t)$  – прогноз накопленного за время  $t$  дебита воды после проведения РИР. Прогноз дебита воды осуществлялся на использовании данных дебитов воды после проведения ремонтно-изоляционных работ до момента сбора данных либо до проведения вторичных ПГИ.

$Q_n(t_{пгИ})'$  – прогноз накопленного дебита нефти на момент проведения первичного ПГИ. Прогноз дебита нефти осуществлялся на использовании данных о проведении промысловых геофизических исследований если бы подрядчик не останавливал скважину на проведение первичного ПГИ.

$Q_n(t_{пгИ2})'$  – прогноз накопленного дебита нефти на момент повторного проведения ПГИ. Прогноз дебита нефти осуществлялся на использовании данных о повторных проведениях промысловых геофизических исследований если бы подрядчик не останавливал скважину на проведение повторных ПГИ.

$Q_n(t_{пгИ'})'$  – прогноз накопленного дебита нефти после РИР на момент проведения повторного ПГИ. Прогноз дебита нефти осуществлялся на использовании данных о времени проведения повторных промысловых геофизических исследованиях после проведения ремонтно-изоляционных работ если бы подрядчик не останавливал скважину на момент проведения повторного ПГИ.

$Q_n(t_{пгИ''})'$  – прогноз накопленного дебита нефти на момент второго проведения ПГИ. Прогноз дебита нефти осуществлялся на использовании данных о времени проведения вторых промысловых геофизических исследований, если бы подрядчик не останавливал скважину на момент проведения второго промыслового геофизического исследования.

$Q_n(t_{пир})'$  – прогноз накопленного дебита нефти на момент проведения РИР. Прогноз дебита нефти осуществлялся на использовании данных о времени проведения ремонтно-изоляционных работ, если бы подрядчик не останавливал скважину на момент проведения ремонтно-изоляционных работ.

$Q_n(t_{rip2})'$  – прогноз накопленного дебита нефти на момент проведения повторного РИР. Прогноз дебита нефти осуществлялся на использовании данных о времени проведения повторного ремонтно-изоляционных работ, если бы подрядчик не останавливал скважину на момент проведения повторных ремонтно-изоляционных работ.

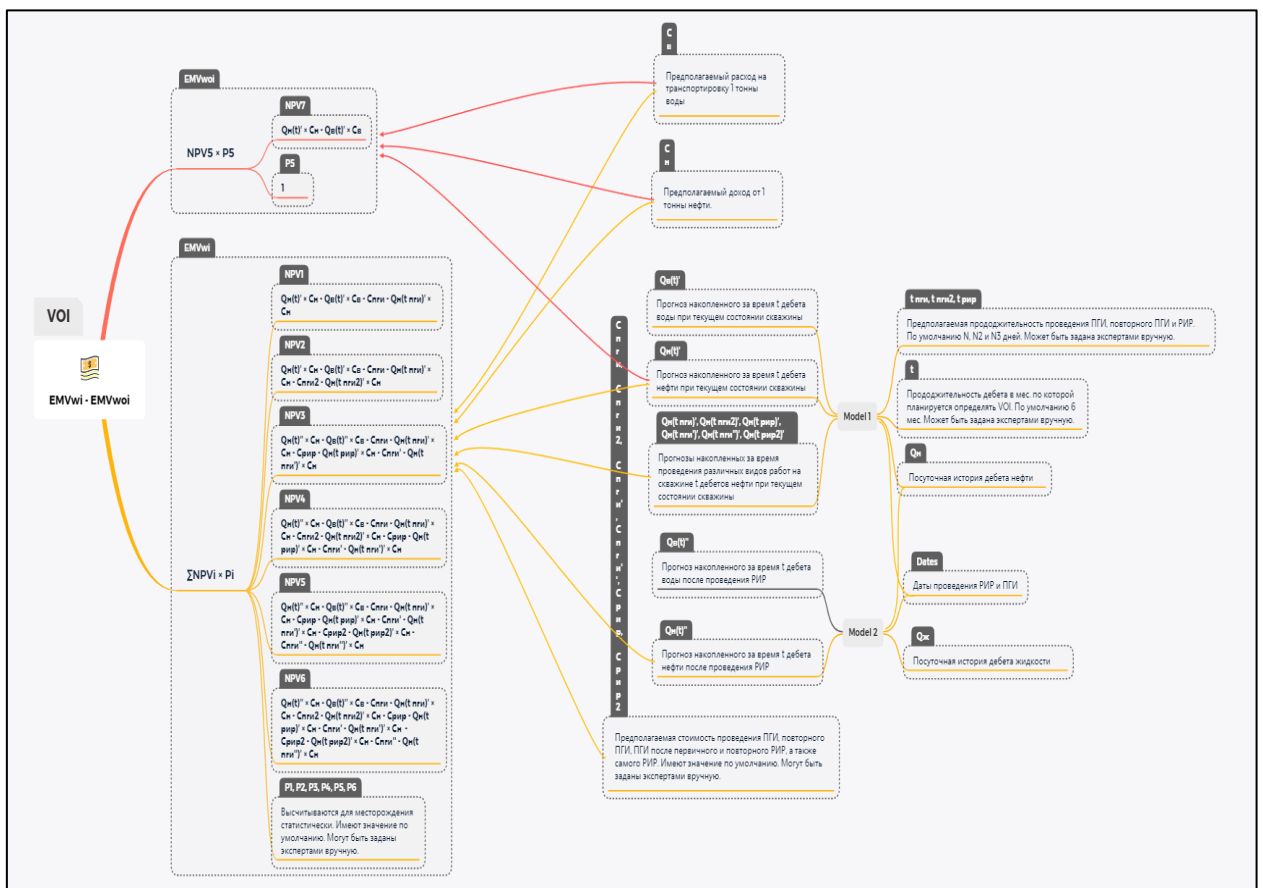


Рис.2.1 Блок-схема расчета VOI для задачи оценки экономической целесообразности проведения ПГИ для РИР на основе исторических данных о скважине.

Исходя из проведенного предварительного анализа исходных данных и возможных вариантов развития событий различных комбинаций промысловых геофизических исследований и ремонтно-изоляционных работ представим в графическом виде в формате блок-схемы математическую

постановку задачи оценки экономической целесообразности проведения ПГИ для РИР на основе исторических данных о скважине.

Прежде чем перейди к непосредственному производству разработки моделей предсказаний дебитов нефти и дебитов воды, проанализируем работу наиболее популярных алгоритмов машинного обучения. Рассмотрим их математические постановки задач. Наилучшие алгоритмы будут реализованы в расчетном программном модуле.

## **2.3 Математические постановки работы алгоритмов машинного обучения для задачи прогнозирования данных.**

### **2.3.1 Математическая постановка алгоритма градиентного бустинга.**

Решается задача восстановления функции в общем контексте обучения с учителем. Будет выбираться набор пар признаков  $\mathcal{X}$  и целевых переменных  $y$ ,  $\{(x_i, y_i)\}_{i=1, \dots, n}$ , на которой будет восстанавливаться зависимость вида  $y = f(x)$ . Восстанавливать будем приближением  $\hat{f}(x)$ , а для понимания также будет функция потерь  $L(y, f)$  которая будет минимизироваться:

$$y \approx \hat{f}(x) \tag{2.14}$$

$$\hat{f}(x) = \arg \min_{f(x)} L(y, f(x)) \tag{2.15}$$

Пока не делаем каких-либо предположений ни о типе зависимости  $f(x)$ , ни о модели наших приближений  $\hat{f}(x)$ , ни о распределении целевой переменной  $y$ . Известно, что функция  $L(y, f)$  должна быть дифференцируемой. Так как задача решается не на всех данных в мире, а только на имеющихся в распоряжении наборе данных, перепишем все в терминах математических

ожиданий. Будем искать приближения функции  $\hat{f}(x)$  так, чтобы в среднем минимизировать функцию потерь на тех данных, что есть:

$$\hat{f}(x) = \arg \min_{f(x)} E_{x,y}[L(y, f(x))] \quad (2.16)$$

Функций  $f(x)$  много так как само их функциональное пространство бесконечномерно. Поэтому чтобы решить задачу, в машинном обучении обычно ограничивают пространство поиска конкретным параметризованным семейством функций  $f(x, \theta), \theta \in \mathbb{R}^d$ . Это упрощает задачу, так как она сводится к уже вполне решаемой оптимизации значений параметров:

$$\hat{f}(x) = f(x, \hat{\theta}) \quad (2.17)$$

$$\hat{\theta} = \arg \min_{\theta} E_{x,y}[L(y, f(\theta))] \quad (2.18)$$

Аналитические решения для получения оптимальных параметров  $\hat{\theta}$  в одну строку существуют достаточно редко, поэтому параметры обычно приближают итеративно. Сначала выписываем эмпирическую функцию потерь  $L_{\theta}(\hat{\theta})$ , показывающую, насколько хорошо произошла оценка по имеющимся у нас данным. После выпишем приближение функции  $\theta$  за  $M$  итераций в виде:

$$\hat{\theta} = \sum_{i=1}^M \hat{\theta}_i \quad (2.19)$$

Осталось подобрать подходящий итеративный алгоритм, которым будет минимизировать функцию потерь  $L_{\theta}(\hat{\theta})$ . Часто используемый вариант — градиентный спуск. Для него выписывается градиент  $\nabla L_{\theta}(\hat{\theta})$  и добавляются наши итеративные оценки  $\hat{\theta}_i$  вдоль него (со знаком минус так как задача уменьшить ошибку, а не нарастить ее). Далее необходимо только инициализировать наше первое приближение оценки  $\hat{\theta}_0$  и выбрать, сколько итераций  $M$  потребуется для завершения процедуры. В неэффективном по

памяти виде хранения приближений  $\hat{\theta}$  алгоритм будет выглядеть следующим образом:

1. Инициализировать начальное приближение параметров  $\hat{\theta} = \hat{\theta}_0$ .
2. Для каждой итерации  $t = 1, \dots, M$  повторять:
  - 2.1 Посчитать градиент функции потерь  $\nabla L_{\theta}(\hat{\theta}) = \left[ \frac{\partial L(y, f(x, \theta))}{\partial \theta} \right]_{\theta = \hat{\theta}}$  при текущем приближении  $\hat{\theta}$ ;
  - 2.2 Задать текущее итеративное приближение  $\hat{\theta}_t$  на основе посчитанного градиента, где  $\hat{\theta}_t \leftarrow \hat{\theta} - \nabla L_{\theta}(\hat{\theta})$ ;
  - 2.3 Обновить приближение параметров  $\hat{\theta}$ , где  $\hat{\theta} \leftarrow \hat{\theta} + \hat{\theta}_t = \sum_{i=1}^t \hat{\theta}_i$ ;
3. Сохранить итоговое приближение  $\hat{\theta}$ , где  $\hat{\theta} = \sum_{i=1}^M \hat{\theta}_i$ ;
4. Пользоваться найденной функцией  $\hat{f}(x) = f(x, \hat{\theta})$  по назначению.

### Функциональный градиентный спуск

Проводим оптимизацию в функциональном пространстве и итеративно ищем приближения функции  $\hat{f}(x)$  в виде самих функций. Выпишем приближение в виде суммы инкрементальных улучшений, каждое из которых является функцией. Для удобства сразу будем считать эту сумму, начиная с начального приближения  $\hat{f}_0(x)$ :

$$\hat{f}(x) = \sum_{i=1}^M f_i(x) \quad (2.20)$$

Далее будем искать наше приближение  $\hat{f}(x)$  не в виде одной модели с множеством параметров (как, например, нейросеть), а в виде суммы функций, чтоб функции продвигались в функциональном пространстве.

Для решения задачи необходимо ограничить поиск функций каким-то семейством функций  $\hat{f}(x) = h(x, \theta)$ . Но сумма моделей может быть сложнее

чем любая модель из этого семейства (сумму двух деревьев-пней глубины 1 уже не приблизить одним пнем). Также общая задача происходит в функциональном пространстве. И на каждом шаге для функций нам понадобится подбирать оптимальный коэффициент  $\rho \in \mathbb{R}$ . Для шага  $t$  задача выглядит следующим образом:

$$\hat{f}(x) = \sum_{i=0}^{t-1} f_i(x) \quad (2.21)$$

$$(\rho_t, \theta_t) = \arg \min_{\rho, \theta} E_{x,y}[L(y, \hat{f}(x) + \rho \cdot h(x, \theta))] \quad (2.22)$$

$$\hat{f}_t(x) = \rho_t \cdot h(x, \theta_t) \quad (2.23)$$

Все наши задачи в общем виде, обучаться на любых моделях  $h(x, \theta)$  при относительно любых функций потерь  $L(y, f(x, \theta))$ . На практике пытаются свести задачу к более простым решениям.

Зная выражение градиента функции потерь, можно посчитать его значения на наших данных. Обучение моделей должно быть наиболее скоррелированными с этим градиентом (со знаком минус). Будем решать задачу МНК-регрессии, с помощью выправления предсказания по этим остаткам. И для классификации, и для регрессии, и для ранжирования в алгоритме все время будет минимизироваться квадрат разности между псевдо-остатками  $r$  и нашими предсказаниями. Для шага  $t$  итоговая задача выглядит следующим образом:

$$\hat{f}(x) = \sum_{i=0}^{t-1} \hat{f}_i(x) \quad (2.24)$$

$$r_{it} = - \left[ \frac{\partial L(y_i, f(x_i))}{\partial f(x_i)} \right]_{f(x)=\hat{f}(x)}, \quad (2.25)$$

где  $i=1, \dots, n$  – число итераций.

$$\theta_t = \arg \min_{\theta} \sum_{i=1}^n (r_{it} - h(x_i, \theta))^2 \quad (2.26)$$

$$\rho_t = \arg \min_{\theta} \sum_{i=1}^n L(y_i, \hat{f}(x_i) + \rho \cdot h(x_i, \theta)) \quad (2.27)$$

### Классический GBM алгоритм Friedman-а

Есть все необходимое, чтобы выписать GBM алгоритм, предложенный Jerome Friedman в 1999 году. Решаем общую задачу обучения с учителем.

На вход алгоритма нужно собрать составляющие такие как:

- набор данных  $\{(x_i, y_i)\}_{i=1, \dots, n}$  ;
- число итераций  $M$  ;
- выбор функции потерь  $L(y, f)$  с выписанным градиентом;
- выбор семейства функций базовых алгоритмов  $h(x, \theta)$ , с процедурой их обучения;
- дополнительные параметры  $h(x, \theta)$ , например, глубина дерева у деревьев решений;

В качестве начального приближения функции  $f_0(x)$  используют просто константное значение  $\gamma$ . Его оптимальный коэффициент  $\rho$  находят бинарным поиском, или другим line search алгоритмом относительно исходной функции потерь (а не градиента). Итак, GBM алгоритм:

1. Инициализировать GBM константным значением  $\hat{f}(x) = \hat{f}_0, \hat{f}_0 = \gamma, \gamma \in \mathbb{R}$  ,

$$\hat{f}_0 = \arg \min_{\gamma} \sum_{i=1}^n L(y_i, \gamma)$$

2. Для каждой итерации  $t = 1, \dots, M$  повторять:

2.1 Посчитать псевдо-остатки  $r_t$

$$r_{it} = - \left[ \frac{\partial L(y_i, f(x_i))}{\partial f(x_i)} \right]_{f(x)=\hat{f}(x)}, \text{ где } i=1, \dots, n$$

2.2 Построить новый базовый алгоритм  $h_t(x)$  как регрессию на псевдо-остатках  $\{(x_i, r_{it})\}_{i=1, \dots, n}$

2.3 Найти оптимальный коэффициент  $\rho_t$  при  $h_t(x)$  относительно исходной

$$\text{функции потерь } \rho_t = \arg \min_{\rho} \sum_{i=1}^n L(y_i, \hat{f}(x_i) + \rho \cdot h(x_i, \theta))$$

2.4 Сохранить  $\hat{f}_t(x) = \rho_t \cdot h_t(x)$

2.5 Обновить текущее приближение  $\hat{f}(x)$ , где  $\hat{f}(x) \leftarrow \hat{f}(x) + \hat{f}_t(x) = \sum_{i=1}^t \hat{f}_i(x)$

3. Скомпоновать итоговую GBM модель  $\hat{f}(x)$ ,  $\hat{f}(x) = \sum_{i=0}^M \hat{f}_i(x)$

### Преимущества и недостатки.

На данный момент алгоритм градиентного спуска является мощных алгоритмом распознавания. Это достигается благодаря адаптивной технике построения композиции. К тому же бустинг предоставляет множество возможностей для вариаций. Во-первых, можно рассматривать различные функции потерь. Это позволяет решать как задачи классификации, так и задачи регрессии. К тому же возможность выбора произвольной функции потерь позволяет акцентировать внимание на особенностях данных в задаче. Во-вторых, возможно рассмотрение любого семейства базовых алгоритмов. А это дает широкие возможности учета особенностей данной задачи. Бустинг над решающими деревьями считается одним из наиболее эффективных вариантов бустинга. Учитывая, что решающие деревья используют базовые алгоритмы (например, пороговые, линейные и т.п.), в результате получается огромное количество вариантов для настройки. В-третьих, благодаря достаточной простоте метода и четкому математическому обоснованию, в каждой конкретной вариации бустинга не сложно провести некоторые математические и алгоритмические оптимизации, которые ускоряют работу алгоритма.

Бустинг не лишен недостатков. Во-первых, бустинг – трудоемкий метод, и работает он достаточно медленно. Часто требуется построение сотен или даже



тысяч базовых алгоритмов для композиции. Во-вторых, без дополнительных модификаций он имеет свойство полностью подстраиваться под данные, в том числе под ошибки и выбросы в них. В-третьих, идея бустинга о плохо применима к построению композиции из достаточно сложных и мощных алгоритмов. Построение такой композиции занимает много времени, а качество существенно не увеличивается. В-четвертых, результаты работы бустинга сложно интерпретируемы, особенно если в композицию входят десятки алгоритмов.

### 2.3.2 Математическая постановка алгоритма Random Forest.

Случайный лес — это ансамблевая модель, состоящая из множества деревьев решений. Прогнозы делаются путем усреднения прогнозов каждого дерева решений. Модель случайного леса также представляет собой набор моделей деревьев решений. Это делает случайные леса сильной техникой моделирования, которая намного мощнее, чем единое дерево решений.

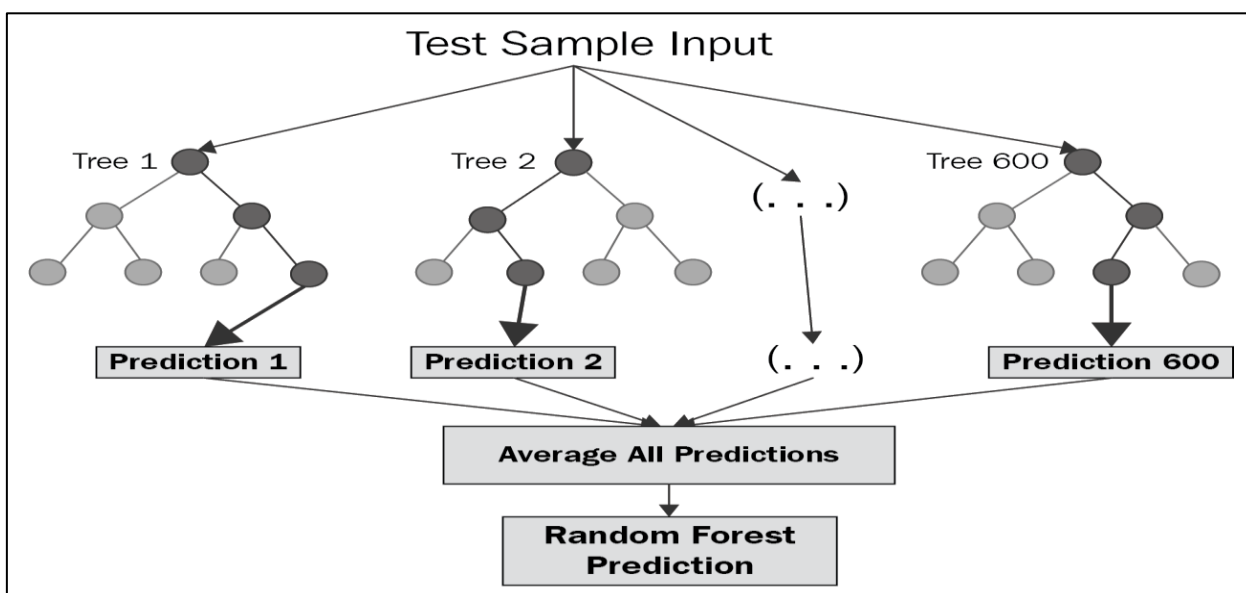


Рис. 2.2 Устройство работы алгоритма Random Forest.

Каждое дерево в случайном лесу обучается на подмножестве предоставленных данных. Подмножество получается как по строкам, так и по столбцам. Это

означает, что каждое случайное дерево леса обучается на случайной выборке точек данных, в то время как в каждом узле принятия решения случайный набор признаков рассматривается для разделения.

В области машинного обучения алгоритм регрессии случайных лесов является наиболее подходящим для задач регрессии, чем другие распространенные и популярные алгоритмы. Ниже приведены случаи, когда алгоритм случайного леса по сравнению с другими алгоритмами регрессии:

1. Существуют нелинейные или сложные отношения между элементами и метками.
2. Вам нужна надежная модель, а это означает, что ее зависимость от шума в тренировочном наборе ограничена. Алгоритм случайного леса является более устойчивым, чем одно дерево решений, так как он использует набор некоррелированных деревьев решений.
3. Если ваши другие реализации линейной модели страдают от переобучения, вы можете использовать случайный лес.

Деревья решений склонны к переоснащению, особенно если не ограничивать максимальную глубину - они обладают неограниченной гибкостью и, могут продолжать расти до тех пор, пока у них не будет ровно один листовой узел для каждой отдельной точки данных, что идеально предсказывает результаты работы алгоритма.

Если ограничить максимальную глубину дерева решений, дисперсия уменьшается, но смещение увеличивается. Чтобы сохранить как дисперсию, так и смещение на низком уровне, алгоритм случайного леса объединяет множество деревьев решений со случайностью для уменьшения переобучения.

Прежде перейти к реализации случайного леса для задачи регрессии, определим некоторые ключевые термины.

Случайный лес — это метаоценщик (т.е. он объединяет результаты нескольких предсказаний), который объединяет множество деревьев решений с некоторыми полезными изменениями:

1. Количество функций, которые можно разделить на каждом узле, ограничено некоторым процентом от общего количества (которое называется гиперпараметром). Это гарантирует, что ансамблевая модель не слишком сильно зависит от какой-либо отдельной функции, и справедливо использует все потенциально прогнозирующие функции.
2. Каждое дерево рисует случайную выборку из исходного набора данных при генерации его расщеплений, добавляя дополнительный элемент случайности, который предотвращает переобучение.

Вышеуказанные модификации помогают предотвратить слишком высокую корреляцию деревьев.

### **Особенности и преимущества случайного леса:**

- Это один из самых точных доступных алгоритмов обучения.
- Он эффективно работает на больших базах данных.
- Он может обрабатывать тысячи входных переменных без удаления переменных.
- Он генерирует внутреннюю непредвзятую оценку ошибки обобщения в процессе строительства леса.

- Он имеет эффективный метод для оценки отсутствующих данных и поддерживает точность, когда отсутствует значительная часть данных.

### **Недостатки случайного леса:**

- Было обнаружено, что случайные леса подходят для некоторых наборов данных с шумными задачами регрессии.
- Для данных, включающих категориальные переменные с различным количеством уровней, случайные леса смещаются в пользу тех атрибутов с большим количеством уровней. Таким образом, значения переменной важности из случайного леса не являются надежными для этого типа данных.

### **2.3.3 Математическая постановка алгоритма KNNRegression.**

Метод ближайших соседей (Англ. *k-nearest neighbors algorithm*, KNN) — метрический алгоритм для автоматической классификации объектов или регрессии.

В случае использования метода для регрессии, объекту присваивается среднее значение по ближайшим к нему объектам, значения которых уже известны. Алгоритм может быть применен к выборкам с большим количеством атрибутов (многомерным). Для этого перед применением нужно определить функцию расстояния; классический вариант такой функции — евклидова метрика. Разные атрибуты могут иметь разный диапазон представленных значений в выборке (например атрибут А представлен в диапазоне от 0.1 до 0,5, а атрибут Б представлен в диапазоне от 1000 до 5000), то значения дистанции могут сильно зависеть от атрибутов с большими диапазонами. Поэтому данные обычно подлежат нормализации. При кластерном анализе есть два основных способа нормализации данных: минимакс-нормализация.

Минимакс-нормализация осуществляется следующим образом:

$$y' = (y - \min[Y]) / (\max[Y] - \min[Y]) \quad (2.28)$$

Где  $\min[Y]$  – минимальное значение ряда;

$\max[Y]$  – максимальное значение ряда;

$y$  – текущее значение ряда;

$y'$  – текущее нормированное значение ряда.

в этом случае все значения будут лежать в диапазоне от 0 до 1.

Уравнение регрессии будет выглядеть следующим образом:

$$\hat{y} = \frac{\sum_{k=1}^K y_k}{K} \quad (2.29)$$

### Взвешенный учет объектов

1. Пусть тренировочная выборка задаётся объектами  $x_1, x_2, \dots, x_K$
2. Упорядочим их относительно рассматриваемого объекта  $\rho(x, x_{i1}) \leq \rho(x, x_{i2}) \leq \dots \rho(x, x_{in})$
3. Обозначим  $z_1 = x_{i1}, z_2 = x_{i2}, \dots, z_K = x_{iK}$
4. Запишем KNN с взвешенным учетом через C-дискриминантную функцию. Для случая с весами вид регрессии будет выглядеть следующим образом:

$$\hat{y} = \frac{\sum_{k=1}^K y_i \cdot \omega(k, \rho(x, z_k))}{\sum_{k=1}^K \omega(k, \rho(x, z_k))} \quad (2.30)$$

Где  $\omega(k, \rho)$  – вес  $k$ -го соседа относительно  $x$ , не возрастает по  $k$ , положителен.

## Достоинства и недостатки.

Выбор числа соседей  $k$ .

При  $k = 1$  алгоритм ближайшего соседа неустойчив к шумовым выбросам: он даёт ошибочные классификации не только на самих объектах-выбросах, но и на ближайших к ним объектах других классов. При  $k = m$ , наоборот, алгоритм чрезмерно устойчив и вырождается в константу. Таким образом, крайние значения  $k$  нежелательны. На практике оптимальное значение параметра  $k$  определяют по критерию скользящего контроля, чаще всего — методом исключения объектов по одному (leave-one-out cross-validation).

Удаление выбросов.

Объекты обучения не являются равноценными, поэтому среди них могут находиться типичные представители классов — *эталоны*. Если классифицируемый объект близок к эталону, то, он принадлежит тому же классу. Ещё одна категория объектов — *неинформативные* или *периферийные*. Они плотно окружены другими объектами того же класса. Если их удалить из выборки, это практически не отразится на качестве классификации. Также в выборку может попасть некоторое количество *шумовых выбросов* — объектов, находящихся «в гуще» чужого класса. Как правило, их удаление только улучшает качество классификации.

Исключение из выборки шумовых и неинформативных объектов даёт несколько преимуществ одновременно: повышается качество классификации, сокращается объём хранимых данных и уменьшается время классификации, затрачиваемое на поиск ближайших эталонов.

Сверхбольшие выборки

Метод ближайших соседей основан на явном хранении всех обучающих объектов. *Сверхбольшие выборки* ( $m \gg 10^3$ ) создают несколько чисто

технических проблем: необходимо не только хранить большой объём данных, но и уметь быстро находить среди них  $k$  ближайших соседей произвольного объекта.

Проблема решается двумя способами:

- выборка *прореживается* путём выбрасывания неинформативных объектов;
- применяются специальные индексы и эффективные структуры данных для быстрого поиска ближайших соседей.

Проблема выбора метрики

В практических задачах классификации редко встречаются такие «идеальные случаи», когда заранее известна хорошая функция расстояния  $\rho(x, x')$ . Если объекты описываются числовыми векторами, часто берут евклидову метрику. Этот выбор, как правило, ничем не обоснован — просто это первое, что приходит в голову. При этом необходимо помнить, что все признаки должны быть измерены «в одном масштабе», а лучше всего — отнормированы. В противном случае признак с наибольшими числовыми значениями будет доминировать в метрике, остальные признаки, фактически, учитываться не будут.

Если признаков слишком много, а расстояние вычисляется как сумма отклонений по отдельным признакам, то возникает проблема «проклятия размерности». Суммы большого числа отклонений с большой вероятностью имеют очень близкие значения (согласно закону больших чисел). Получается, что в пространстве высокой размерности все объекты примерно одинаково далеки друг от друга; выбор  $k$  ближайших соседей становится практически произвольным.

Проблема решается путём отбора относительно небольшого числа информативных признаков. В алгоритмах вычисления оценок строится

множество различных наборов признаков, для каждого строится своя функция близости, затем по всем функциям близости производится голосование.

### 2.3.4 Математическая постановка алгоритма ARIMA.

**ARIMA** (англ. *autoregressive integrated moving average* интегрированная модель авторегрессии–скользящего среднего — модель и методология анализа временных рядов. Является расширением моделей *ARMA* для нестационарных временных рядов, которые можно сделать стационарными взятием разностей некоторого порядка от исходного временного ряда (разностно-стационарные временные ряды). Модель  $ARIMA(p, d, q)$  означает, что разности временного ряда  $d$  подчиняются модели  $ARMA(p, q)$ .

Модель  $ARIMA(p, d, q)$  для нестационарного временного ряда  $u_t$  имеет вид:

Пусть имеется скалярный временной ряд  $\{u_i\}_{i=1}^N$ . Необходимо найти (предсказать) значения величины  $v$  в моменты времени  $t > t_N$ . Регрессия, не подходит для описания большинства временных зависимостей, так как при этом предполагают явный вид функциональной зависимости  $u$  от  $t$ . Выход: моделирование ряда в виде случайного процесса  $u = u(t_i, \omega)$ :

где  $\omega$  — случайное событие (отсюда  $u_i$  — случайная величина);

$p$  — порядок модели  $AR(p)$ ;

$d$  — порядок интегрирования;

$q$  — порядок модели  $MA(q)$ .

Предположим, что базовая моделью нашего временного ряда является



нормальный белый шум  $a(t)$ . Значения в такой модели в различные моменты времени независимы, а значение в любой фиксированный момент времени распределено по нормальному закону:

$$p(a) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-a^2}{2\sigma^2}} \quad (2.31)$$

Где  $a$  – белый шум,  $\sigma$  – дисперсия.

Для такой модели среднее значение принято считать равным нулю, значение автокорреляционной функции при любой задержке (ненулевой) равно нулю. Остается оценить только дисперсию. Чаще всего получается, что нельзя применить модель нормального белого шума к выборке (например, значения автокорреляционной функции отличны от нуля). Тогда будем говорить о процессе как о нормальном белом шуме, преобразованном линейным фильтром:

$$u_n = a_n + \sum_{i=1}^{\infty} \psi_i a_{n-i} \quad (2.32)$$

$$\sum_{i=1}^{\infty} \psi_i^2 \leq const \quad (2.33)$$

Несложно понять, что с увеличением  $i$  значения  $\psi_i$  быстро убывают, следовательно можно ограничиться количеством весов  $q$ .

$$u_n = a_n + \sum_{i=1}^q \theta_i a_{n-i} \quad (2.34)$$

Переписывая выражение нормального белого шума, преобразованного линейным фильтром, приходим к формуле:

$$u_n = a_n + \sum_{i=1}^{\infty} \pi_i u_{n-i} \quad (2.35)$$

Снова понимаем, что бесконечное количество весов нам не нужно, ограничиваемся количеством весов  $p$ :

$$u_n = a_n + \sum_{i=1}^p \phi_i u_{n-i} \quad (2.36)$$

Соединив модель скользящего среднего и модель авторегрессии, получим

модель  $ARMA(p, q)$ :

$$u_n = a_n + \sum_{i=1}^p \phi_i u_{n-i} - \sum_{i=1}^q \theta_i a_{n-i} \quad (2.37)$$

Проблема заключается в том, что если ряд имеет признаки не стационарности, то модель не может быть адекватной. Но стационарной может быть некоторая разность наблюдаемого процесса порядка  $d$  :

$$\omega_n = \nabla^d u_n \quad (2.38)$$

$$\nabla u_n = u_n - u_{n-1} \quad (2.39)$$

Переходим к  $ARIMA(p, q, d)$ :

$$\omega_n = a_n + \sum_{i=1}^p \phi_i u_{n-i} - \sum_{i=1}^q \theta_i a_{n-i} + \mu \quad (2.40)$$

Ограничения:

1. Считаем наблюдаемые процессы случайными. Отсюда следует, что нельзя использовать динамические модели. Вместо этого будут использоваться статистические модели и аппараты статистики и теории вероятностей.
2.  $ARIMA$  — это пример эмпирической модели, конструируемой на экспериментальных данных.
3.  $ARIMA$  использует структурные методы анализа временного ряда. Эти методы, в отличие от неструктурных, ориентированы на оценку по небольшому количеству параметров при дополнительных предположениях о свойствах наблюдаемых величин.

### 2.3.5 Математическая постановка для алгоритма $SARIMA$ .

Для построения авторегрессионной модели на временных рядах с присутствием сезонного компонента используется расширение очередное расширение модели  $ARIMA$  —  $SARIMA$ , где  $S$  означает seasonal (сезонность). В общем виде модель выглядит так:  $ARIMA(p, d, q)(P, D, Q)_S$

где  $p, d, q$  — старые параметры пришедшие из  $ARIMA$ ,  $(P, D, Q)_S$  — группа новых параметров, специфичных для  $SARIMA$ . Они задают, насколько сезонов мы смотрим назад, когда её предсказываем. Рассмотрим каждый параметр подробнее.

$p$  — порядок модели  $AR(p)$ ;

$d$  — порядок интегрирования;

$q$  — порядок модели  $MA(q)$ ;

$P$  — порядок сезонной составляющей  $SAR(P)$ ;

$D$  — порядок интегрирования сезонной составляющей;

$Q$  — порядок сезонной составляющей  $SMA(Q)$ ;

$S$  — размерность сезонности.

### **Преимущества и недостатки моделей $ARIMA/SARIMA$ .**

К преимуществам можно отнести то, что эти модели имеют очень четкое математико-статистическое обоснование, что делает их одними из наиболее научно обоснованных моделей из всего множества моделей прогнозирования тенденций во временных рядах.

Еще одно преимущество является формализованная и наиболее подробно разработанная методика, следуя которой можно подобрать модель, наиболее подходящую к каждому конкретному временному ряду. Формальная процедура проверки модели на адекватность достаточно проста.

Также точечные и интервальные прогнозы следуют из самой модели и не требуют отдельного оценивания.

Один из явных недостатков моделей заключается в требовании к рядам данных: для построения адекватной модели *ARIMA* требуется не менее 40 наблюдений, а для *SARIMA* – порядка 6–10 сезонов, что на практике не всегда возможно.

Вторым серьезным недостатком является не адаптивность моделей авторегрессии: при получении новых данных модель нужно периодически переоценивать, а иногда – и переидентифицировать.

Третий недостаток заключается в том, что построение удовлетворительной модели *ARIMA* требует больших затрат ресурсов и времени. Требуется большой опыт со стороны прогнозиста.

Но эти все преимущества и недостатки касаются лишь процесса построения модели.

Все построение моделей *ARIMA* основывается на предположении о том, что временной ряд генерируется бесконечно в соответствии с какой-то функцией, параметры которой нам нужно идентифицировать и оценить, т. е. в основе подхода *ARIMA* лежит предположение о застывшем характере протекающих процессов, эволюционность как таковая в модели не учитывается. Вызвано это не в последнюю очередь тем, что модели изначально разрабатывались для моделирования физических и технических

У моделей *ARIMA* есть свои недостатки, лежащие в самой их основе. Однако это ни в коей мере не говорит о том, что от этих моделей надо отказаться. Для каждого конкретного случая стоит обращаться к своей прогнозной модели: будь то простейшие модели, модели трендов, сезонной декомпозиции, модели экспоненциального сглаживания или модели авторегрессий со скользящей средней. Просто стоит иметь в виду как положительные, так и отрицательные

стороны используемых моделей и опираться на те прогнозы, относительно которых можно сказать, что они лучше опишут реальную ситуацию в будущем.

Из рассмотрения математических постановок задач машинного обучения вытекает новая задача для работы с этими алгоритмами. Чтоб качественно работали алгоритмы машинного обучения, данные, которые поступают на расчет необходимо предобработать. Во входных данных необходимо устранить недочеты: восстановить пропуски значений исследуемых величин, удалить выбросы и дубликаты значений, для более качественной работы алгоритмов. В следующем параграфе будет показана предобработка входных данных со скважин для дальнейшего их использования в алгоритмах машинного обучения.

#### **2.4 Предварительная обработка входных данных для прогнозирования расчетных моделей для дебитов нефти и воды.**

В процессе своей деятельности нефтяные компании получают большое количество данных, которые необходимо обрабатывать и анализировать для повышения эффективности работы предприятий. Для этой цели используются современные технологии обработки больших данных, которые предоставляют различные инструменты анализа и предсказания будущих трендов в области геологии и нефтегазового производства.

Куст	71	1314			зобой иск. 2871,30			Пласт	кол. добыча нефти	интервал перфорации	Перфоратор	Дата посл. перф.	Удл.	Газ. фактор	Рпл	НКТ	ГРП			КВН (м/л)	Уьем проп.																			
		Цех	ЦДНГ-7		Б88/1	105321,96	5,90	2825;2825 -	К-105;ЛКС-8	09.11.2003	106	136,59	17,985	62/2268-252	05.02.2003	5894,413.00.2004																								
ЗУ	Дата	Режимы	Параметр	Средн.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31					
Воронка, глуб. 2800 м., МРП, ос. ос.	Май 2019		Ф вращ ТМ																																					
			Обв																																					
			Обв ХАЛ	85,3																																				
			Обв	76,75	71	59	70	78	80	78	80	63	68	78	80	80	79	73	68	74	80	76	81	81	77	81	78	81	83	83	82	81	84	80	80					
			Ф																																					
			Дшт																																					
			Обв ТМ (иск)	76,75	71	59	70	78	80	78	80	63	68	78	80	80	79	73	68	74	80	76	81	81	77	81	78	81	83	83	82	81	84	80	80					
			Рбуф ТМ																																					
			Qr (АГЗУ)																																					
			Ртр																																					
			Рбуф ТМ																																					
			ГФ																																					
			Обв ТМ ручн																																					
			Обв ТМ ручн																																					
			Разб(РтрТМ)																																					
			Рват ТМ																																					
			Примечание																																					

Рис.2.3 Образец входных данных по «шахматкам».

Предприятием был предоставлен набор данных по 30 добывающим скважинам из месторождения X и заключения ПГИ для этих скважин. Они включают в себя посуточные дебиты нефти, посуточные дебиты воды, даты проведения ГРП, данные об обводненности скважин и заключения проведенных ПГИ и РИР, и других вспомогательных мероприятий.

Исполнитель ПГИ				ПЦН		Заказчик ПГИ	
Месторождение				Скважина	Куст	Дата исследования	
2314.4				2316.4	00534	20/03/2017	
2389.2				2390.6			
Заключение ПГИ							
Определение технического состояния скважин							
Способ эксплуатации при исследовании - АЗОПИРОВАНИЕ							
Приборы							
Тип прибора	Номер	Пороговое значение					
КСАТ-8	690						
Конструкция скважины							
Диаметр эксплуатационной колонны:	178 мм						
Диаметр НКТ:	73 мм						
Диаметр хвостовика:	114 мм						
Искусственный забой:	3334 м						
Глубина забоя:	2441 м						
Интервал измерительного твара:	2314,4-2316,4м; 2389,2-2390,6 м						
Глубина спуска НКТ по данным промысла:	1700 м						
Результаты измерения давления и уровней							
Дата измерения	Время	Давление, атм	Г.Д.Р., м	В.Н.Р., м			
20/03/2017	16.06		917	946			
	16.17		900	950			
	16.38		873	925			
	17.02		851	901			
	17.27		827	880			
	20.47		778	830			
Данные по перфорации							
Пласт	Интервал проволочной, м	Интервал по Г.Д.Р., м	Тип перфорации		Дата перфорации		
БВВ-1	3102.38 - 3103.08						
БВВ-1	3226.11 - 3226.81						
БВВ-1	3323.5 - 3333.68						
Техническое состояние скважины							
Кровля, м	Подштан, м	Состояние скважины					

Барометрич.				
Режимы	Дата	Время	Давление, атм	Глубина, м
Фондовый замер	20/03/2017	02:00	160.2	2300
Фондовый замер	20/03/2017	06:00	157.16	2300
При компрессировании	20/03/2017	07:00	181.68	2300
При компрессировании	20/03/2017	08:40	187.15	2300
При компрессировании	20/03/2017	09:10	185.41	2300
После прогона пусковой муфты	20/03/2017	09:30	176.8	2300
После прогона пусковой муфты	20/03/2017	09:50	172.22	2300
После прогона пусковой муфты	20/03/2017	10:20	157.64	2300
После прогона пусковой муфты	20/03/2017	11:00	145.3	2300
После прогона пусковой муфты	20/03/2017	12:00	139.67	2300
После прогона пусковой муфты	20/03/2017	13:00	138.94	2300
После прогона пусковой муфты	20/03/2017	14:06	136.57	2300
После прогона пусковой муфты	20/03/2017	14:30	136.4	2300
При стабилизации	20/03/2017	14:38	82.47	2300
При стабилизации	20/03/2017	14:39	83.31	2300
При стабилизации	20/03/2017	14:40	84.46	2300
При стабилизации	20/03/2017	14:41	85.52	2300
При стабилизации	20/03/2017	14:42	86.53	2300
При стабилизации	20/03/2017	14:43	87.51	2300
При стабилизации	20/03/2017	14:44	88.46	2300
При стабилизации	20/03/2017	14:45	89.42	2300
При стабилизации	20/03/2017	14:53	96.45	2300
При стабилизации	20/03/2017	15:28	119.03	2300
При стабилизации	20/03/2017	15:33	121.19	2300
При стабилизации	20/03/2017	15:50	127.1	2300
При стабилизации	20/03/2017	18:00	141.38	2300
При стабилизации	20/03/2017	18:20	142.06	2300
При стабилизации	20/03/2017	18:40	142.72	2300
При стабилизации	20/03/2017	19:00	143.28	2300
При стабилизации	20/03/2017	19:20	143.77	2300
При стабилизации	20/03/2017	19:40	144.21	2300
При стабилизации	20/03/2017	20:05	144.75	2300
При стабилизации	20/03/2017	20:28	145.11	2300

Дополнение:  
Выполнены детальные исследования в интервале 2140-2441,6м. Статистический уровень на глубине 549м. На фоне интенсивного всходящего потока с глубины 2441,6м "голова"

Рис.2.4 Образец заключения ПГИ.

Перед тем, чтобы преступить непосредственно к прогнозированию дебитов нефти и воды, для работы алгоритмов машинного обучения необходимо произвести предварительную обработку данных. Она в себя включает:

нормирование значений, восстановление пропусков значений, удаление выбросов и удаление дубликатов значений исследуемого набора данных.

Подготовка и чистка исследуемого набора данных для дальнейшей визуализации производилась на языках программирования Python. Также были использованы следующие инструменты и библиотеки:

1. NumPy – это библиотека языка Python, добавляющая поддержку больших многомерных массивов и матриц, вместе с большой библиотекой высокоуровневых (и очень быстрых) математических функций для операций с этими массивами.
2. Pandas – программная библиотека на языке Python для обработки и анализа данных. Работа pandas с данными строится поверх библиотеки NumPy, являющейся инструментом более низкого уровня.
3. Seaborn – это библиотека визуализации данных Python, основанная на Matplotlib. Она предоставляет высокоуровневый интерфейс для рисования информативной статистической графики.

Совершим очистку данных. Очистка данных – это процесс обнаружения и исправления или удаления поврежденных или неточных записей из набора записей, таблицы или базы данных. В результате выявляются неполные, неправильные, неточные или несущественные данных, а после замену, изменение или удаление выбросов.

До очистки некачественных данных, всегда нужно представлять исходный набор данных (см. Рис. 3. и Рис.4.).

В «шахматках» представлены данные примерно с 1985 по 2015 ( $\pm 3$  года). И в таблицах есть множество пропусков и параметром для нашего исследования.

Нет общих решений для проблемы восстановления отсутствующих данных. Для каждого конкретного набора данных приходится искать наиболее подходящие методы или их комбинации.

Используем две самые распространенные техники предобработки данных:

## 1. Отбрасывание записей.

В нашем наборе данных примерно с 1985 по 2011 ( $\pm 2$ года) отсутствуют значения у исследуемых признаков и так же во время ПГИ и РИР происходила остановка скважин, замеры не производились, поэтому строки с отсутствующими значениями можно удалить.

## 2. Отбрасывание признаков.

Для нашего исследования нам необходимы данные по суточному дебиту нефти ( $Q_n$ ) и суточному дебиту жидкости ( $Q_{ж}$ ). Остальные признаки для исследования не вносят информативность и их удалим.

В сформированном нами НД остались пропуски значений (вследствие улучшения технологий, более устаревшие данные задокументированы хуже). Затем формировался новый НД. Из первоначального НД производили разбиение суточных дебитов нефти и жидкости по неделям, методом экстраполяции значений создавали новый набор данных дебитов нефти и жидкости разделенные на недельные циклы.

Для дальнейшего использования набора данных необходимо произвести нормирование значений данных. Эта манипуляция производится для дальнейшего использования в алгоритмах машинного обучения. Нормирование данных необходимо для того, чтобы выраженность некоторого качества описать числом. Обычно  $x$  меняется от некоторого минимального значения  $x_{\min}$  (отражающего отсутствие качества) до некоторого максимального значения  $x_{\max}$  (крайняя степень проявления, наличия, выраженности). Его получение решает проблему сравнения двух объектов, но только по этому показателю.

На основании требований к предварительной обработке данных был разработан отдельный расчетный модуль для автоматизации этой задачи.



В следующем параграфе будет подробно продемонстрирована разработанная методология расчета параметров ожидаемых ценностей проекта, при наличии дополнительной информации и без ее отсутствия (состояние «как есть»).

## **2.5 Методика расчета показателей ожидаемых ценностей проекта, при наличии дополнительной информации и при отсутствии дополнительной информации.**

В параграфе 2.2 была составлена в общем виде блок-схема методологии расчета экономической ценности информации на основе исторических данных со скважины. Проанализируем более подробно её основные составляющие.

Её можно разделить на 2 глобальные части:

1. Методология расчета показателя ожидаемой ценности информации при наличии дополнительной информации
2. Методология расчета показателя ожидаемой ценности информации при отсутствии дополнительной информации

### **2.5.1 Алгоритм расчета показателя ожидаемой ценности информации при наличии дополнительной информации.**

Представим в виде блок-схемы расчет показателя ожидаемой ценности информации при наличии дополнительной информации ( $EMV_{wi}$ ).

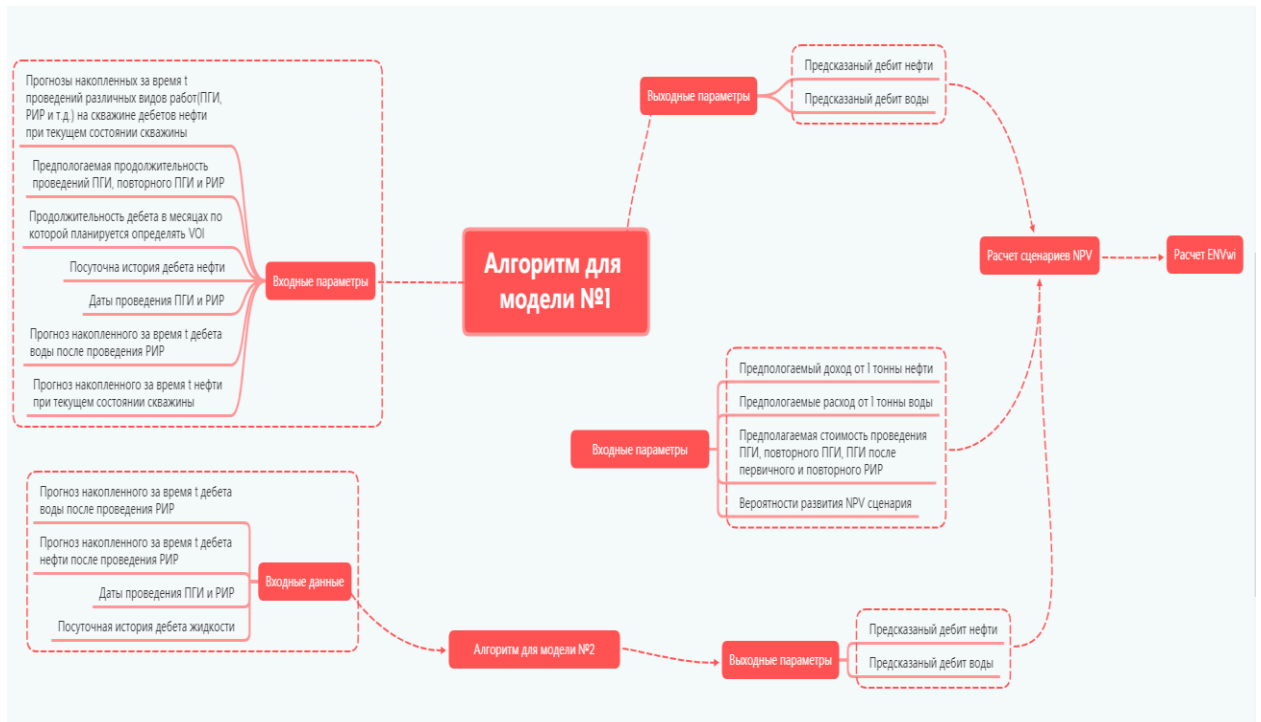


Рис.2.5 Упрощенная блок-схема расчета показателя ожидаемой ценности информации с учетом дополнительной информации ( $EMV_{wi}$ ).

Расчетная модель включает в себя два алгоритма расчета. Алгоритмы отличаются в основном набором входных параметров. Для первого алгоритма использовались данные до проведения ПГИ и РИР.

Входными параметрами являются алгоритма 1:

- прогнозы накопленной нефти за время проведения ПГИ, РИР; продолжительность времени ПГИ, РИР, повторных ПГИ;
- продолжительность дебетов в месяцах, по которым планируется рассчитывать VOI;
- даты проведения ПГИ и РИР;
- прогноз накопленного дебита воды за время проведения ПГИ и РИР;
- прогноз накопленного дебита нефти за время проведения ПГИ и РИР.

Выходные параметры:

- прогноз дебита нефти;
- прогноз дебита воды.

По умолчанию прогнозы дебитов строятся на 6 месяцев (количество месяцев, на которые строятся прогнозы указываются специалистом).

Входные данные для алгоритма 2:

- посуточная история дебета жидкости;
- даты проведения ПГИ и РИР;
- прогноз дебита нефти после проведения ПГИ и РИР;
- прогноз накопленного дебита воды после проведения ПГИ и РИР.

Выходные данные работы алгоритма 2:

- прогноз дебита нефти;
- прогноз дебита воды.

По умолчанию прогнозы дебитов строятся на 6 месяцев (количество месяцев, на которые строятся прогнозы указываются специалистом).

После поступления входных параметров их обрабатывают предсказательные алгоритмы машинного обучения. В качестве предсказательных алгоритмов расчета дебитов нефти и жидкости использовались несколько алгоритмов машинного обучения. Их более детальное рассмотрение будет сделано в главе 3. Лучшие алгоритмы будут реализованы в программном расчетном модуле. После того как произошло предсказание, следует расчет сценариев денежных потоков (NPV). Для сценариев NPV с дополнительной информацией было получено 6 возможных исходов событий (см. глава 2, пар.2.2). И для

получения значения ожидаемой ценности проекта с учетом дополнительной информации ( $ENV_{wi}$ ) необходимо выполнить произведения соответствующей вероятности сценария на сумму значений платежей с учетом дополнительной информации ( $NPV1-NPV6$ ) и посчитать их сумму. Вероятности развития сценария оцениваются экспертом и высчитываются при помощи статистического анализа. В следующем параграфе рассмотрим ожидаемой ценности информации при отсутствии дополнительной информации.

## 2.5.2 Алгоритм расчета показателя ожидаемой ценности информации при отсутствии дополнительной информации.

Представим в виде блок-схемы расчет показателя ожидаемой ценности информации при отсутствии дополнительной информации ( $EMV_{woi}$ ).

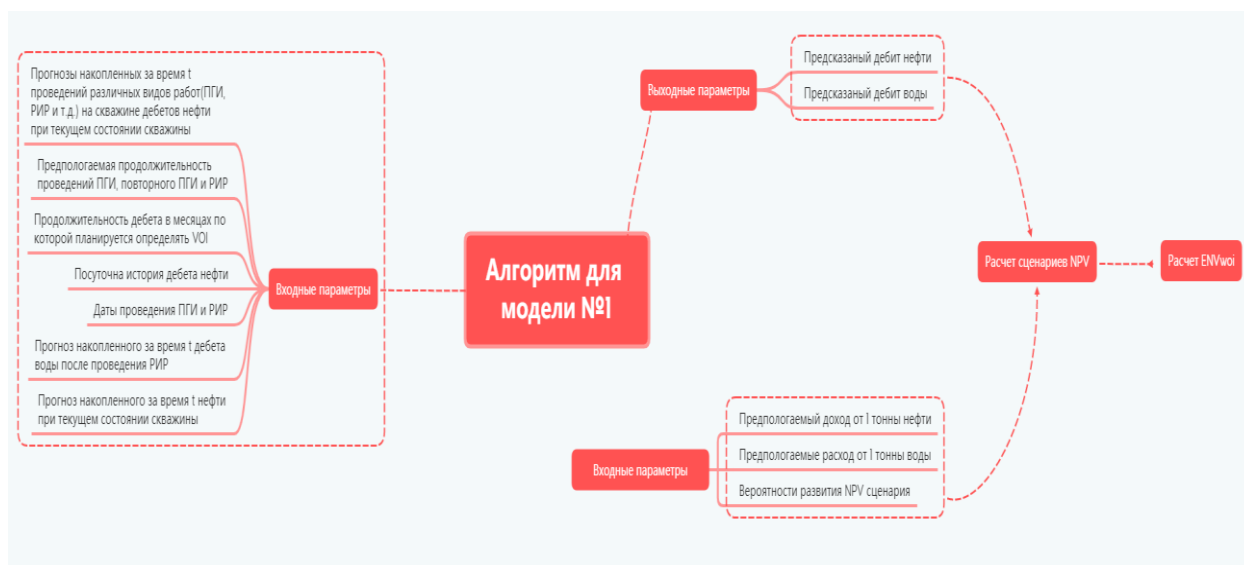


Рис.2.6 Упрощенная блок-схема расчета показателя ожидаемой ценности информации без учета дополнительной информации ( $EMV_{woi}$ ).

Для первого алгоритма использовались данные до проведения ПГИ и РИР. Входными параметрами являются алгоритма 1:

- прогнозы накопленной нефти за время проведения ПГИ, РИР; продолжительность времени ПГИ, РИР, повторных ПГИ;
- продолжительность дебетов в месяцах, по которым планируется рассчитывать VOI;
- даты проведения ПГИ и РИР;
- прогноз накопленного дебита воды за время проведения ПГИ и РИР;
- прогноз накопленного дебита нефти за время проведения ПГИ и РИР.

Выходные данные:

- прогноз дебита нефти;
- прогноз дебита воды.

По умолчанию прогнозы дебитов строятся на 6 месяцев (количество месяцев, на которые строятся прогнозы указываются специалистом).

После поступления входных параметров их обрабатывают предсказательные алгоритмы машинного обучения. В качестве предсказательных алгоритмов расчета дебитов нефти и жидкости использовались несколько алгоритмов машинного обучения. Их более детальное рассмотрение будет сделано в главе 3. После того как произошло предсказание, следует расчет сценариев денежных потоков без учета дополнительной информации (NPV). Для сценариев NPV при отсутствии дополнительной информацией был получен 1 возможный исход сценария (см. глава 2, пар.2.2). И для получения значения ожидаемой ценности проекта с учетом дополнительной информации ( $ENV_{woi}$ ) необходимо выполнить произведения соответствующей вероятности сценария на сумму значений платежей при отсутствии дополнительной информации скважины ( $NPV1 - NPV6$ ) и посчитать их сумму. Вероятности развития сценария оцениваются экспертом и высчитываются при помощи статистического анализа.

## 2.6 Выводы к главе 2.

В этой главе был проведен анализ общей структуры расчета VOI для оценки экономической целесообразности проведения промысловых геофизических исследований скважин для ремонтно-изоляционных работ. Были получены все возможные сценарии расчетов приведённых денежных потоков к сегодняшнему дню, с учетом и без учета дополнительной информации. Получилось 6 вариантов развития событий расчетов дебетов нефти и воды с учетом дополнительной информации о скважинах и 1 вариант расчета дебета нефти и дебета воды, без учета информации. Разработали блок-схемы расчетов ожидаемой ценности проекта с учетом дополнительной информации и ожидаемой ценности проекта без учета дополнительной информации. Был проведен анализ различных алгоритмов машинного обучения и были определены их преимущества и недостатки. Графическая реализация окончательной расчетной блок-схемы для расчета VOI для нашего исследования и численная реализация алгоритмов будет получена в следующей главе.

### **ГЛАВА 3. ТЕХНИЧЕСКАЯ РЕАЛИЗАЦИИ МЕТОДОЛОГИИ VOI ДЛЯ ЗАДАЧИ ОЦЕНКИ ЭКОНОМИЧЕСКОЙ ЦЕЛЕСООБРАЗНОСТИ ПРОВЕДЕНИЯ ПРОМЫСЛОВЫХ ГЕОФИЗИЧЕСКИХ ИССЛЕДОВАНИЙ ДЛЯ РЕМОНТНО- ИЗОЛЯЦИОННЫХ РАБОТ НА ОСНОВЕ ИСТОРИЧЕСКИХ ДАННЫХ О СКВАЖИНЕ.**

#### **3.1 Предварительная обработка входных данных для прогнозирования параметров расчетных моделей для дебитов нефти и воды.**

Подготовка и чистка исследуемого набора данных для дальнейшей визуализации производилась на языках программирования Python. Также были использованы следующие инструменты и библиотеки:

NumPy – это библиотека языка Python, добавляющая поддержку больших многомерных массивов и матриц, вместе с большой библиотекой высокоуровневых (и очень быстрых) математических функций для операций с этими массивами.

Pandas – программная библиотека на языке Python для обработки и анализа данных. Работа pandas с данными строится поверх библиотеки NumPy, являющейся инструментом более низкого уровня.

Seaborn – это библиотека визуализации данных Python, основанная на Matplotlib. Она предоставляет высокоуровневый интерфейс для рисования информативной статистической графики.

SciPy — библиотека в составе Python SciPy Stack, не запутайтесь. Она содержит модули для линейной алгебры, оптимизации, интеграции и статистики. SciPy работает совместно с NumPy, что позволяет ей значительно расширить функциональность. Два аргумента, почему эта библиотека действительно хороша:



- Наличие подпрограмм, таких как численная интеграция и оптимизация, для повышения производительности.
- Подробная документация на каждую функцию.

Библиотека Prophet, разработана для прогнозирования большого числа различных бизнес-показателей и строит достаточно хорошие модели прогнозов данных. Кроме того, библиотека дает возможность, изменяя человеко-понятные параметры, улучшать прогноз и не требует от аналитиков глубоких знаний устройства предсказательных моделей.

На основе описанных выше библиотек был написан дополнительный расчетный модуль для автоматизации процесса предварительной обработки входных данных с «шахматок» и заключений промысловых геофизических исследований для добывающих скважин.

Опишем алгоритм работы данного программного модуля. Вспомогательный модуль для извлечения данных из экселевских файлов с «шахматками» нефтяных скважин по ссылке на корневую папку или на файл. Из файлов извлекаются среднесуточные значения по месяцам и ежедневные значения, проводится анализ динамики дебита нефти и воды, показателя обводненности, расчет NPV при проведении РИР/ПГИ и без проведения ремонтных работ.

Если изменение дебита за последние 2 года превышает критический уровень, для скважины производится расчет NPV по сценариям ремонта. Достижимый эффект от ремонта (увеличение дебита нефти, снижение дебита воды) определяется по наилучшему значению за последние 2 года с поправкой в 5%. Извлеченные из файлов и обработанные данные с прогнозами сохраняются в файл 'Candidates.xlsx'.

Расшифровка наименований столбцов в итоговом наборе данных:

Q\_oil – средний дебит нефти за последний месяц, т в сутки;

Max\_oil – максимальный дебит нефти в предшествующий период, т в сутки в среднем за 1 месяц;

Oil\_delta – изменение дебита нефти в долях от 1 ( $Q_{oil} / Max_{oil}$ );

Oil\_total – накопленная добыча нефти, т;

Q\_water – средний дебит воды за последний месяц, т в сутки;

Min\_water – минимальный дебит воды в предшествующий период, т в сутки в среднем за 1 месяц;

Water\_delta – изменение дебита воды в долях от 1 ( $Q_{water} / Min_{water}$ );

Water\_pc – средняя обводненность за последний месяц, %;

Min\_water\_p – минимальный процент обводненности в предшествующий период;

Water\_pc\_delta – изменение обводненности в долях от 1 ( $Water_{pc} / Min_{water_{pc}}$ );

fbprophet\_oil – прогноз дебита суммарного нефти за прогнозный период;

fbprophet\_water – прогноз дебита суммарного воды за прогнозный период;

NPV\_as\_is – ожидаемый доход (убыток) при отсутствии ремонтных работ;

NPV\_repair – ожидаемый доход (убыток) от проведения ремонтных работ;

Scenario\_1 <...> Scenario\_6 – ожидаемый доход (убыток) по сценариям;

Last\_perf – дата последней перфорации;

G' – дата ГРП;

Defect – наличие повреждения скважины (True - вероятно, False - не обнаружено);

Defect\_date – дата возникновения дефекта (определяется пользователем по графику);

Repair – рекомендации по ремонту (True - ремонт рекомендуется, False - не рекомендуется);

Path – путь к файлу с «шахматкой» скважины.

### **Алгоритм предварительной обработки входных данных.**

1. Вывод информации о том, запускается скрипт самостоятельно или импортируется из другого расчётного модуля.
2. Визуализация входных данных, представленных в формате экселевских таблиц, и настройка параметров отображения этих таблиц.

3. Ввод параметров отображения графиков, входных таблиц:
4. Составление словаря для преобразования текстовых названий месяцев из исследуемых таблиц в числовой формат данных.
5. Нахождение алгоритмом корневой папки, нахождение в ней всех экселевских файлов, содержащих слово «Шахматка» и алгоритм, возвращает список путей к этим файлам.
6. Создание функции, преобразующей строку «Месяц гггг» в объект типа Period.
7. Функция преобразует строку «дд Месяц гггг» в объект типа datetime.
8. Функция принимает путь к экселевскому файлу с «шахматкой» скважины, извлекает среднесуточные данные.
9. Возвращает набор данных с параметрами скважины и Period-индексом ежемесячной частоты.
10. Выполнение преобразования данных из строкового типа данных в числовой тип данных:
11. Создаем параметры с заголовками столбцов, а периоды - индексами строк:
12. Преобразуем значения столбца в Period:
13. Удаляем столбцы, где нет ни одного значения:
14. Добавляем столбец с содержанием воды и рассчитываем его значение:
15. Создаем функцию, принимающую путь к экселевскому файлу с «шахматкой» скважины и извлекаем ежедневные данные из неё.
16. Возвращаем набор данных с параметрами скважины и datetime-индексом ежедневной частоты.
17. Извлекаются данные  $Q_n$ ,  $Q_{ж}$ , Обводненность, «Примечание». Примечания преобразуются в бинарный числовой столбец, где 0 – отсутствие примечаний, 1 – наличие примечаний.
18. Далее производим интерполяцию данных в пустых ячейках.
19. Заполняем даты в объединённых ячейках.

- 20.Выбираем строки, содержащие необходимые параметры для дальнейшего прогнозирования.
- 21.Для ускорения обработки данных оставляем только строки за последние 10 лет:
- 22.Удаляем пропуски в столбце с датой (несуществующие дни месяца):
- 23.Преобразуем значения параметров в числовой тип данных:
- 24.Делаем параметры заголовками столбцов, а даты – индексами строк:
- 25.Добавляем столбец с содержанием воды и рассчитываем его значение.
- 26.Создаем функцию, принимающую набор данных «monthly\_df» с ежемесячными данными по скважине. После функция извлекает средние значения дебита нефти и воды, среднюю обводненность за последний месяц и возвращает полученные значения.
- 27.Если данные по обводненности не заполнены, рассчитываем обводненность.
- 28.Функция принимает набор данных «monthly\_df» с ежемесячными данными по скважине, словарь «cur\_vals» со сводными показателями. Параметр «period» определяет продолжительность периода в месяцах, за который производится расчет изменения дебита от пиковых значений к текущему среднему.
- 29.Определяем пиковые значения и изменение дебита нефти и воды, показателя обводненности за указанный период (по умолчанию - за 2 года) в долях от 1.
- 30.Возвращаем дополненный словарь «cur\_vals» со сводными показателями.
- 31.Сокращаем временной ряд в соответствии с «period» и отбрасываем последний месяц с данными.
- 32.Вычисляем максимальный дебит нефти и добавляем в словарь.
- 33.Вычисляем минимальный дебит воды и обводненность, добавляем в словарь:

34. Создаем функцию принимающую строку с данными, идентифицирующую номер скважины.
35. Возвращаем числовое значение, отсекая буквы в конце строки.
36. Функция принимает путь к корневой папке, находит в ней по ключевым словам
37. html-файлы с данными ПГИ, возвращает словарь, содержащий списки путей к файлам по ключам - номерам скважин. Номер скважины извлекается из названия папки нижнего уровня.
38. Словарь для добавления путей к html-файлам по номерам скважин:
39. Ищем ключевые слова в названиях html-файлов:
40. Извлекаем номер скважины из названия папки и добавляем путь в словарь:
41. Убираем буквенные символы.
42. Функция принимает список путей к html-файлам с данными ПГИ.
43. Алгоритм обрабатывает файлы и возвращает список дат ПГИ по скважине.
44. Извлекаем дату из ячейки таблицы.
45. Функция принимает набор данных «agg\_df» со сводными данными по скважинам, критические коэффициенты изменения дебита нефти и воды, сравнивает фактические изменение дебита с критическими значениями, добавляет в столбец «Defect» значение True, если дефект скважины вероятен, False – если изменение дебита не критично.
46. Возвращает дополненный набор данных со сводными данными по скважинам.
47. Функция принимает аргумент «period», определяющий кол-во месяцев для расчета NPV, набор данных «agg\_df» со сводными данными по скважинам, словарь параметров материальных и временных затрат и вероятностей по 6-ти сценариям «pars».
48. «oil\_loss» и «water\_increase» – коэффициенты, отражающие достижимый после ремонта уровень дебита от максимума по нефти и

минимума по воде за предшествующие 2 года с учетом поправки на естественные процессы.

49. Функция рассчитывает значения NPV по 6-ти сценариям ремонтных работ и средний ожидаемый эффект, NPV без проведения ремонта. Возвращает дополненный набор данных.
50. Заполняем значения денежных потоков (NPV) без ремонта (умножаем прогнозируемый суммарный дебит нефти и воды на удельную стоимость):
51. Выбираем скважины с вероятным дефектом и заполняем значения NPV по сценариям.
52. Вычитаем из всего периода длительность ремонтных работ, умножаем на дебит и удельные показатели по воде и нефти, вычитаем суммарные затраты по каждому из 6-ти сценариев.
53. Перемножаем значения денежных потоков (NPV) по сценариям на вероятности того какой из сценариев произойдет.
54. Обновляем значения в основном наборе данных.
55. Функция принимает набор данных с ежедневными данными по скважине и период в месяцах. Если в наборе данных есть данные за более ранние периоды, чем указанный, убирает их. Возвращает набор данных за указанный период.
56. Функция принимает набор данных со среднемесячными данными по скважине и путь к файлу, выводит график среднесуточного дебита нефти и воды за последние 3 года эксплуатации скважины.
57. Функция принимает набор данных с ежедневными данными по скважине, путь к файлу и необязательный параметр window (по умолчанию - 7 дней).
58. Вычисляет скользящее среднее значение для указанного «окна».

59. Применяет метод «convolve» библиотеки «scipy» для идентификации моментов выраженной смены тренда в рядах данных. Выводит графики исходных значений дебита нефти и воды и изменений сигнала в данных.
60. Выводим график с исходными данными и сигналом:
61. Функция принимает набор данных с ежедневными данными по скважине, прогнозный период (количество месяцев) и наименование параметра (столбца набора данных).
62. Выполняет прогноз дебита этого параметра в указанный период с применением модели «fbprophet».
63. Возвращает прогнозируемый суммарный дебит.
64. Извлечение из «шахматок» данных для модели.
65. Инициализация и обучение модели.
66. Создание временной шкалы для прогнозируемого периода.
67. Выполнение прогноза.
68. Расчет суммарного дебита в прогнозный период.
69. Отрицательные прогнозы преобразуются в ноль.

### **3.2 Техническая реализация расчетных алгоритмов машинного обучения для получения параметров ожидаемых ценностей проекта с учетом дополнительной информации и ожидаемых ценностей проекта без учета дополнительной информации.**


В главе 2 были рассмотрены способы преобразования входных данных для дальнейшей их работы в алгоритмах машинного обучения, были рассмотрены математические постановки задач для реализуемых алгоритмов машинного обучения и были разработаны алгоритмы в виде блок-схем для расчетов параметров ожидаемых ценностей проектов с учетом дополнительной информации и без учета дополнительной информации о скважинах. В этом параграфе будет определено какие из выбранных алгоритмов машинного обучения будут наиболее оптимально предсказывать результаты исследуемых параметров.

### 3.2.1 Техническая реализация расчетных алгоритмов машинного обучения для получения параметров ожидаемых ценностей проекта с учетом дополнительной информации.

Рассмотрим расчет параметра ожидаемой ценности проекта с учетом дополнительной информации. При его расчете ключевыми параметрами которые оказывают влияние на конечный результат расчета  $NPV$  и как следствие и  $ENV_{wi}$  – это предсказанный дебит нефти и предсказанный дебит воды. Предварительный анализ алгоритмов машинного обучения (параграфы 2.3.1–2.3.3) показал, что три алгоритма показывают хорошие предсказательные результаты для задачи прогнозов дебитов нефти и воды с учетом дополнительной информации. Это алгоритм градиентного бустинга и алгоритмы KNNRegression и Random Forest. Все эти алгоритмы успешно используются в анализе данных. Рассмотрим техническую реализацию каждого из алгоритмов на примере скважины 1448.

#### 3.2.1.a Техническая реализация алгоритма Random Forest, для задачи прогнозирования дебитов нефти и воды.

Проанализируем работу алгоритма Random Forest. Алгоритм расчета представлен на рисунке 12.



```

RF = RandomForestRegressor(random_state=15)

params = {"max_depth": [5, 40],
          "n_estimators": [20, 50, 100, 300, 500],
          "max_features": [3, 17],
          "min_samples_split": [2, 3, 10],
          "min_samples_leaf": [1, 3, 10],
          "bootstrap": [True, False],
          "criterion": ["mse"]}

grid_search = GridSearchCV(RF, param_grid=params, cv=3, n_jobs=-1, verbose=True)
grid_search.fit(X_train, y_train)

grid_search.best_params_
best_regr = RandomForestRegressor(max_depth=None, bootstrap=False, max_features=5, min_samples_leaf=10, min_samples_split=2, n_estimators=30, random_state=
best_regr.fit(X_train, y_train)

RandomForestRegressor(bootstrap=False, ccp_alpha=0.0, criterion='mse',
max_depth=None, max_features=5, max_leaf_nodes=None,
max_samples=None, min_impurity_decrease=0.0,
min_impurity_split=None, min_samples_leaf=10,
min_samples_split=2, min_weight_fraction_leaf=0.0,
n_estimators=30, n_jobs=None, oob_score=False,
random_state=10, verbose=0, warm_start=False)

```

Рис.3.1 Техническая реализация алгоритма Random Forest для скважины 1448, реализованная на языке Python.



В состав алгоритма входят множество параметров. Для исследования алгоритма важны 7 параметров, остальные это надстройки не сильно оказывающие внимание на окончательный результат работы алгоритма. Проанализируем какую функцию исполняет каждый из этих параметров.

`max_depth` – Максимальная глубина дерева. Если `None`, то узлы деревьев расширяются до тех пор, пока все листья не станут чистыми или пока все листья не будут содержать менее чем минимальное количество выборок, необходимое для разделения внутреннего узла.

`n_estimators` – Количество деревьев в лесу.

`max_features` – Количество функций, которые следует учитывать при поиске лучшего разделения:

- Если `int`, то особенности рассматриваются параметра `max_features` при каждом разделении.
- Если с плавающей точкой, то параметр `max_features` является дробью, и особенности рассматриваются при каждом разделении.
- Если «авто», то `max_features=n_features`.
- Если «sqrt», то `max_features=sqrt(n_features)`.
- Если «log2», то `max_features=log2(n_features)`.
- Если нет, тогда `max_features=n_features`.

`min_samples_split` – Минимальное количество выборок, необходимое для разделения внутреннего узла:

`min_samples_leaf` – Минимальное количество образцов которое должно быть в листовом узле. Точка разделения на любой глубине будет рассматриваться только в том случае, если она оставляет хотя бы минимальное количество образцов в листовом узле обучающей выборки в каждой из левой и правой ветви. Это может оказать эффект сглаживания модели, это важно особенно в моделях регрессии.

`bootstrap` – Используются ли образцы начальной загрузки при построении деревьев. Если `False`, весь набор данных используется для построения каждого дерева.

`criterion` – Функция для измерения качества разделения деревьев. Поддерживаемыми критериями являются «mse» для среднеквадратичной ошибки, которая равна уменьшению дисперсии в качестве критерия выбора признаков, и «mae» для средней абсолютной ошибки.

Для прогноза алгоритма были выбраны параметры, изображенные на рисунке 6 выделенные красным цветом. Они использовались для проверки работы алгоритма. Сам же оптимальный подбор параметров реализовывался через функцию `Grid_search` из библиотеки `scikit-learn`. Функция выбрала следующие оптимальные параметры:

`max_depth` – `None` (узлы деревьев расширяются до тех пор, пока все листья не станут чистыми или пока все листья не будут содержать менее чем минимальное количество выборок, необходимое для разделения внутреннего узла).

`n_estimators` – 30 – Количество деревьев в лесу.

`max_features` – 5 – количество функций, которые следует учитывать при поиске лучшего разделения.

`min_samples_split` – 2 – минимальное количество выборок, необходимое для разделения внутреннего узла

`min_samples_leaf` – 10 – минимальное количество образцов которое должно быть в листовом узле.

`bootstrap` – `False` – весь набор данных используется для построения каждого дерева.

`criterion` – `mae` – критерий средней абсолютной ошибки.

Средняя абсолютная ошибка  $MAE = 716,775$

### 3.2.1.6 Техническая реализация алгоритма KNNRegression, для задачи прогнозирования дебитов нефти и воды.

Проанализируем какую функцию исполняет каждый из этих параметров.

Цель прогнозируется локальной интерполяцией целей, связанных с ближайшими соседями в обучающем наборе, следующими параметрами:

`n_neighbors` – количество соседей.

`weights` – весовая функция, используемая в прогнозировании. Возможные значения:

«uniform»: единообразные веса. Все точки в каждой окрестности имеют одинаковый вес.

«distance»: вес указывает на обратную величину их расстояния. в этом случае более близкие соседи точки запроса будут иметь большее влияние, чем соседи, которые находятся дальше.

«callable»: пользовательская функция, которая принимает массив расстояний и возвращает массив той же формы, содержащий веса.

Унифицированные веса используются по умолчанию.

`algorithm` – алгоритм, используемый для вычисления ближайших соседей:

'ball\_tree' будет использовать BallTree

'kd\_tree' будет использовать `scipy.spatial.cKDtree`

'brute' будет использовать поиск грубой силы.

'auto' будет пытаться выбрать наиболее подходящий алгоритм на основе значений, переданных для метода соответствия .

Примечание: установка на разреженный вход переопределит настройку этого параметра, используя грубую силу.

leaf\_size – размер листа передается BallTree или cKDTree. Это может повлиять на скорость построения и запроса, а также на объем памяти, необходимый для хранения дерева. Оптимальное значение зависит от характера проблемы.

В данном исследовании наименьшую ошибку показали данные по умолчанию, выглядят они следующим образом:

n\_neighbors=5 – количество соседей.

weights='uniform' – весовая функция, используемая в прогнозировании. «uniform»: единообразные веса. Все точки в каждой окрестности имеют одинаковый вес.

algorithm='auto' – алгоритм, используемый для вычисления ближайших соседей «auto» будет пытаться выбрать наиболее подходящий алгоритм на основе значений, переданных для метода соответствия .

leaf\_size=30 – размер листа передается BallTree или cKDTree. Это может повлиять на скорость построения и запроса, а также на объем памяти, необходимый для хранения дерева.

Средняя абсолютная ошибка MAE = 182,2

### **3.2.1.в Техническая реализация алгоритма градиентного бустинга, для задачи прогнозирования дебитов нефти и воды.**

Перед запуском XGBoost мы должны установить три типа параметров: общие параметры, параметры бустера и параметры задачи.

- Общие параметры относятся к тому, какой бустер мы используем для бустинга, обычно это дерево или линейная модель.
- Параметры бустера зависят от того, какой бустер вы выбрали

- Параметры учебного задания зависят от учебного сценария. Например, задачи регрессии могут использовать разные параметры с задачами ранжирования.

### **Общие параметры**

`Booster` – какой усилитель использовать. Может быть `gbtree`, `gblinear` или `dart`; `gbtree` и `dart` использовать модели на основе «деревьев», в то время как `gblinear` использует линейные функции.

`verbosity` – многословие печати сообщений. Допустимые значения: 0 (молчание), 1 (предупреждение), 2 (информация), 3 (отладка). Иногда XGBoost пытается изменить конфигурации на основе эвристики, которая отображается в виде предупреждающего сообщения.

`validate` – при значении `True` XGBoost выполнит проверку входных параметров, чтобы проверить, используется ли параметр или нет.

`nthread` – количество параллельных потоков, используемых для запуска XGBoost, по умолчанию максимальное количество доступных потоков, если не установлено.

`disable_default_eval_metric` – флаг, чтобы отключить метрику по умолчанию.

`num_rbuffer` – размер буфера прогнозирования, обычно равный числу обучающих экземпляров. Буферы используются для сохранения результатов прогнозирования последнего шага повышения. Устанавливается автоматически XGBoost.

`num_feature` – размер элемента, используемый в бустинге, установлен на максимальный размер элемента. Устанавливается автоматически XGBoost.

## Параметры для Tree Booster :

`eta` – размер шага, используемый в обновлении, предотвращает переобучение. После каждого шага повышения есть возможность напрямую получать веса новых функций и `eta` сокращает веса функций, чтобы сделать процесс повышения более консервативным.

`Gamma` – минимальное уменьшение потерь, необходимое для создания дальнейшего разбиения на листовом узле дерева. Чем больше `gamma`, тем более консервативным будет алгоритм.

`max_depth` – максимальная глубина дерева. Увеличение этого значения сделает модель более сложной. Ноль принимается только в `lossguided` растущей политике, когда `tree_method` установлен как, `hist` и это указывает на отсутствие ограничений по глубине. Помните, что `XGBoost` активно потребляет память при обучении глубокому дереву.

`min_child_weight` – минимальная сумма веса экземпляра (гессiana), необходимая обучающей выборке. Если в результате шага разбиения дерева конечный узел `min_child_weight` получит сумму веса экземпляра меньше, чем, процесс сборки прекратит дальнейшее разбиение. В задаче линейной регрессии это просто соответствует минимальному количеству экземпляров, которое должно быть в каждом узле. Чем больше `min_child_weight`, тем более консервативным будет алгоритм.

`max_delta_step` – максимальный шаг дельты мы разрешаем каждому листу выводить. Если значение установлено 0, это означает, что ограничений нет. Если для него установлено положительное значение, это может помочь сделать шаг обновления более консервативным. Обычно этот параметр не нужен, но он может помочь в логистической регрессии, когда класс крайне несбалансирован.

`subsample` – соотношение подвыборок учебных экземпляров. Установка его на 0,5 означает, что `XGBoost` будет случайным образом отбирать половину

обучающих данных до выращивания деревьев и это предотвратит переобучение. Соотношение подвыборок будет происходить один раз в каждой итерации повышения.

`sampling_method` – метод, используемый для выборки учебных экземпляров.

`uniform`: каждый обучающий экземпляр имеет равную вероятность быть выбранным. Обычно устанавливается `subsample> = 0,5` для хороших результатов.

`gradient_based`: вероятность выбора для каждого обучающего экземпляра пропорциональна *нормализованному абсолютному значению* градиентов. `Subsample` может быть установлен на уровне 0,1 без потери точности модели.

`colsample_bytree`, `colsample_bylevel`, `colsample_bynode` – это семейство параметров для подвыборки столбцов.

`colsample_bytree` – отношение выборки столбцов при построении каждого дерева. Подвыборка выполняется один раз для каждого построенного дерева.

`colsample_bylevel` – коэффициент подвыборки столбцов для каждого уровня. Подвыборка выполняется один раз для каждого нового уровня глубины, достигнутого в дереве. Столбцы подвергаются выборке из набора столбцов, выбранных для текущего дерева.

`colsample_bynode` – отношение выборки столбцов для каждого узла (разделение). Соотношение подвыборок происходят один раз каждый раз, когда оценивается новый сплит. Столбцы подвергаются выборке из набора столбцов, выбранных для текущего уровня.

`Lambda` – член регуляризации L2 по весам. Увеличение этого значения сделает модель более консервативной.

`Alpha` – член регуляризации L1 по весам. Увеличение этого значения сделает модель более консервативной.

`scale_pos_weight` – параметр контролирует баланс положительных и отрицательных весов, полезных для несбалансированных классов. Типичное значение для рассмотрения.

Строка, разделенная запятыми, определяющая последовательность запуска средств обновления дерева, обеспечивая модульный способ построения и изменения деревьев. Это расширенный параметр, который обычно устанавливается автоматически, в зависимости от некоторых других параметров. Тем не менее, он также может быть установлен явно пользователем. Существуют следующие средства обновления:

`grow_colmaker` – нераспределенное построение деревьев на основе столбцов.

`grow_histmaker` – построение распределенного дерева с разбивкой данных на основе строк на основе глобального предложения подсчета гистограмм.

`grow_local_histmaker` – на основе локального подсчета гистограммы.

`grow_skmaker` – использует приблизительный алгоритм зарисовки.

`grow_quantile_histmaker` – вырастить дерево с использованием квантованной гистограммы.

`grow_gpu_hist` – вырастить дерево с помощью графического процессора

`sync` – синхронизирует деревья во всех распределенных узлах.

`Refresh` – обновляет статистику дерева и / или значения листьев на основе текущих данных. Обратите внимание, что случайная подвыборка строк данных не выполняется.

`refresh_leaf` – это параметр средства `refresh` обновления. Когда этот флаг равен 1, обновляются как листья дерева, так и статистика узлов дерева. Когда это 0, только статистика узла обновляется.

`process_type` – тип ускорения процесса запуска.



Выбор: default, update

Default – нормальный процесс повышения, который создает новые деревья.

Update – запускается из существующей модели и обновляет только ее деревья. На каждой итерации повышения берется дерево из исходной модели, для этого дерева запускается определенная последовательность обновлений, и к новой модели добавляется измененное дерево. Новая модель будет иметь одинаковое или меньшее количество деревьев, в зависимости от количества выполненных форсирующих итераций.

Predictor – тип используемого алгоритма предсказания. Предоставляет те же результаты, но позволяет использовать графический процессор или процессор.

Auto – Настройка предсказателя на основе эвристики.

cpu\_predictor – Многоядерный алгоритм прогнозирования процессора.

gpu\_predictor – Прогнозирование с использованием графического процессора.

num\_parallel\_tree – количество параллельных деревьев, построенных во время каждой итерации.

Использование алгоритма XGB с параметрами по умолчанию дало очень хороший прогноз. Параметры по умолчанию выглядят следующим образом:

#### **Параметры по умолчанию:**

booster = gbtrees

verbosity = 1

validate\_parameters = true

disable\_default\_eval\_metric = 0

#### **Параметры по умолчанию для Tree Booster:**

eta = 0,3

gamma = 0

max\_depth = 6

min\_child\_weight = 1

max\_delta\_step = 0

subsample = 1

sampling\_method = uniform

lambda = 1

alpha = 0

sketch\_eps = 0.03

scale\_pos\_weight = 1

process\_type = default

max\_leaves = 0

num\_parallel\_tree = 1

Средняя абсолютная ошибка MAE = 141,383

Таблица №1 «Расчет средней абсолютной ошибки и средней абсолютной ошибки, выраженной в процентах.

Well	mae(KNN)	mape(KNN)	mae(RF)	mape(RF)	mae(XGB)	mape(XGB)
1001	444,2666667	16,0477773	996,975	36,0127	128,844	4,654094694
1175	1612,7	39,7902788	2080,86	51,3412	944,3086	23,29900305
1314	142,9333333	13,8905086	215,31	20,9242	104,844	10,18892516
1357	246,7	23,1447603	133,17	12,4937	620,5419	58,21764662
1448	182,2	46,010101	716,775	181,004	141,383	35,70278341
1704	547,3333333	27,9095066	27,54	1,40431	559,9406	28,55237167
172P	375,7	92,5597438	552,165	136,035	127,3429	31,37297197
1730	363,8333333	27,7460027	642,15	48,9705	663,5735	50,60424805
1820	71,96666667	2,41994239	380,175	12,7837	282,6822	9,505437172
1919	61,06666667	4,97083164	772,23	62,8596	635,4583	51,72635592
1930	249,2666667	96,3908224	474,99	183,677	118,7161	45,90723321
2019	1044,6	56,7563162	1044,6	56,7563	849,4952	46,15567556
2071	18,7	2,57150715	108,78	14,9587	125,6177	17,27415878
2115	190,6	61,2074502	238,845	76,7004	83,63034	26,856244
2575	4595,7	64,0533534	4552,91	63,4569	5860,728	81,68489769
39	478,1	18,5482619	292,11	11,3326	1000,391	38,81093279
4202	1376,2	72,641858	736,44	38,8725	1995,082	105,3091452
4206	115,1333333	17,3498091	545,715	82,2355	268,5502	40,46868726
424P	29,26666667	2,23443783	479,46	36,6056	280,6938	21,43028039
4309	536,6	1688,66667	1148,24	3727,45	383,6615	1178,871632
5150	1107,366667	30,775573	1815,68	50,4606	753,0798	20,92934692
5301	656,6333333	21,7572344	766,485	25,3971	92,26369	3,057113556
534	640,2666667	34,4284921	33,15	1,78255	543,3016	29,21447385
570	380,3	75,8173844	849,855	169,429	662,1534	132,008245
628	272,0333333	54,4938568	447,615	89,6665	44,08618	8,83136598
6372	2246,3	220,875123	2476,7	243,529	1964,838	193,1993873

Необходимо определить какой алгоритм дает наименьшую ошибку между реальными данными и предсказательными данными. На основании данных, представленных в таблице 1, построим график относительной ошибки для всех трех алгоритмов для каждой скважины (см. рис.6).

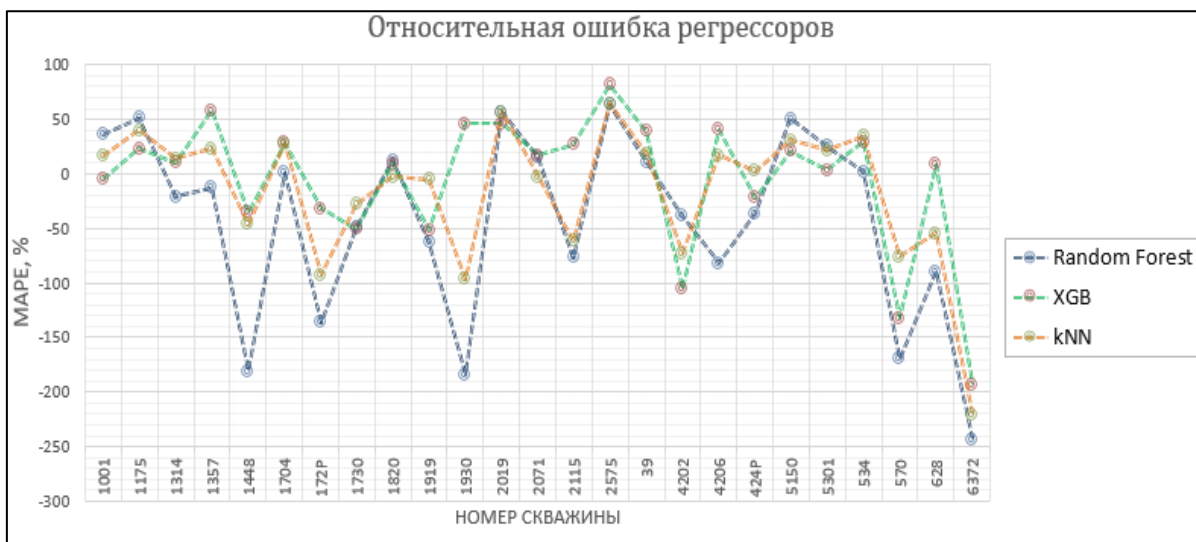


Рис.3.2 Сравнение относительной ошибки предсказательных алгоритмов для каждой исследуемой скважины.

Из графика сравнений относительной ошибки предсказательных алгоритмов для каждой из исследуемой скважины видно, что наименьшую относительную ошибку предсказательного алгоритма получается при использовании алгоритма градиентного бустинга. Именно он будет реализован в программном расчетном модуле.

Для случаев предсказания дебитов нефти и дебитов воды с учетом дополнительной информации получается 6 сценариев расчета денежного потока, приведенного к сегодняшнему дню с учета дополнительной информации ( $NPV1 - NPV6$ ). Поэтому расчет  $ENV_{wi}$  для каждого сценария сводится к простому произведению денежных потоков, с учетом дополнительной информации на вероятности сценариев, которые могут произойти.

### 3.2.2 Техническая реализация расчетных алгоритмов машинного обучения для получения параметров ожидаемых ценностей проекта при отсутствии дополнительной информации.

Рассмотрим расчет дебитов нефти и дебитов жидкости без учета дополнительной информации. Для решения этих задач рассматривались алгоритмы ARIMA и SARIMA из библиотеки Stat.models (параграфы 2.3.4 и 2.3.5) и для каждой скважины происходил отдельный расчет алгоритма. Подбор для всех скважин параметров  $p, d, q$  осуществлялся методом перебора через функцию grid search из библиотеки scikit-learn. Данные, на которых происходило обучение были предшествующие ПГИ и РИР. Временной ряд делился на полугодовые промежутки. Дальше для каждого промежутка считалась относительная ошибка. В конечном счете получали суммарную ошибку методом усреднения по участкам.

#### Рассмотрим работу алгоритма ARIMA на примере скважины 1448.

1. Найдем в данных даты проведения ПГИ, РИР или ГРП. Визуализируем наш временной ряд с учетом дат проведения мероприятий (см. рис.7).

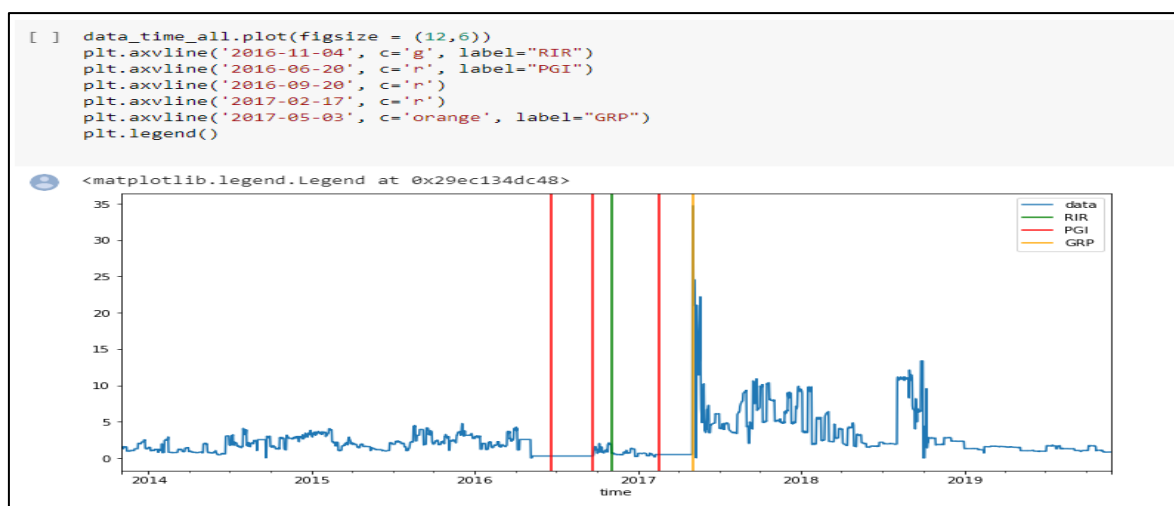


Рис.3.3 Визуализация набора данных в виде временного ряда для скважины 1448.

Для первого шестимесячного промежутка параметры алгоритма ARIMA

получились следующими (см. рис.9):

$p$  – порядок модели  $AR(p) = 2$

$d$  – порядок интегрирования = 0

$q$  – порядок модели  $MA(q) = 2$

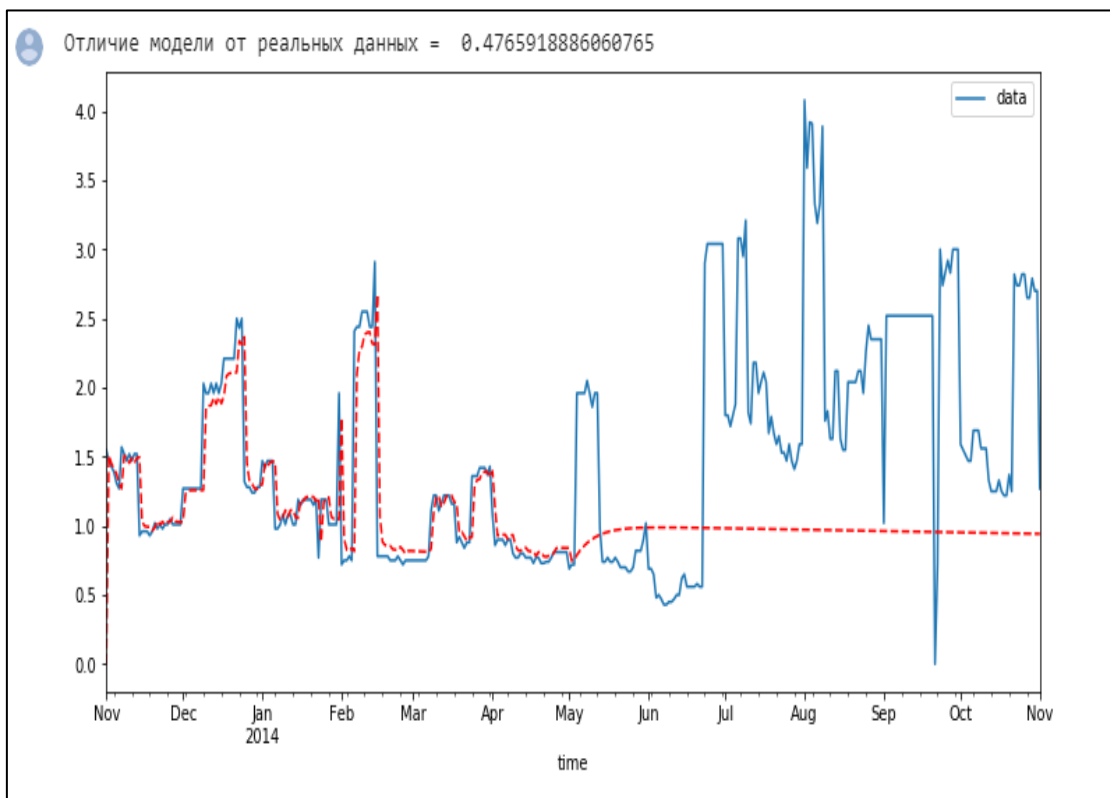


Рис.3.4 Прогноз дебита нефти скважины 1448 в первый шестимесячный промежуток.

Эти параметры были подобраны функцией подбора оптимальных параметров `grid_search` из библиотеки `sklearn`, для остальных полугодовых промежутков были проведены аналогичные алгоритмы действий. В конце методом осреднения по участкам считалась средняя абсолютная ошибка,  $MAE = 0,2475$ .

```
[ ] index_start_train = data_time_all.index[0]
    index_finish_train = data_time_all.index[0] + timedelta(6 * 365 / 12)
    index_start_pred = data_time_all.index[0] + timedelta(6 * 365 / 12)
    index_finish_pred = data_time_all.index[0] + timedelta(12 * 365 / 12)

    data_time_train = data_time_all[index_start_train:index_finish_train]
    model_arima = sm.tsa.SARIMAX(data_time_train, order=(2, 0, 2), freq = 'D').fit(full_output = False, disp = 0)
    pred_arima = model_arima.predict(index_start_train, index_finish_pred)

    data_time_all[index_start_train:index_finish_pred].plot(figsize = (12,6))
    pred_arima.plot(style = 'r--')
    ## Считаем сумму за предсказанный период
    sum_pred = pred_arima[index_start_pred:index_finish_pred].sum()
    sum_real = data_time_all[index_start_pred:index_finish_pred].sum()
    delta = (sum_real - sum_pred) / sum_real
    print("Отличие модели от реальных данных = ", delta.data)
```

Рис 3.5 Программная реализация алгоритма ARIMA для скважины 1448 в первый шестимесячный промежуток, реализованная на языке Python.

Для случаев предсказания дебитов нефти и дебитов воды без учета дополнительной информации у нас только один сценарий расчета денежного потока, приведенного к сегодняшнему дню без учета дополнительной информации ( $NPV_7$ ). Поэтому расчет  $ENV_{woi}$  сводится к простому произведению денежного потока, без учета дополнительной информации на вероятность того, что событие произойдет. Аналогично процедура продельвается для каждой скважины. В программном расчетном модуле для расчета денежной стоимости проекта без учета дополнительной информации для предсказания дебитов жидкости и нефти будет реализовываться предсказательный алгоритм ARIMA.

### 3.2 Программная реализация расчетного модуля VOI для задачи оценки экономической целесообразности проведения промышленных геофизических исследований для ремонтно-изоляционных работ на основе исторических данных о скважине.

#### Краткое описание пользовательского интерфейса.

При запуске программы появляется интерактивное окно, в котором пользователь указывает ссылку на корневую папку, содержащую файлы с данными по скважинам, и вводит необходимые для работы скрипта параметры

Кандидаты РИР

Шаг 1. Нажмите 'Выбрать папку' и в открывшемся окне выберите нужную директорию.  
 Шаг 2. Введите материальные и временные затраты: целые числа без пробелов и знаков препинания.  
 Шаг 3. Введите вероятности сценариев ремонта: целые числа от 0 до 100 без пробелов и знаков препинания.  
 ПОЯСНЕНИЕ. Сумма вероятностей по заполненным сценариям должна составлять 100.  
 Если один или несколько сценариев не рассматриваются, оставьте соответствующие поля пустыми.  
 Если все сценарии равновероятны, оставьте все поля пустыми.  
 Шаг 4. Введите критические значения падения и роста дебита для периода в 2 года: целые положительные числа без пробелов и знаков препинания.  
 ПРИМЕР 1. Введите 20 в ячейку 'Δ нефти', если критическим падением дебита следует считать 20%.  
 ПРИМЕР 2. Введите 30 в ячейку 'Δ воды', если критическим ростом дебита следует считать 30%.  
 Шаг 5. Нажмите 'Запустить программу'.

C:/Users/dranitsyna/Desktop/Al\_folder

Заполнить все поля		Заполнить, если вероятности не одинаковы		Описание сценариев при проведении ремонтных работ:
Руб./т нефти	<input type="text"/>	Сценарий 1	<input type="text"/>	
Руб./т воды	<input type="text"/>	Сценарий 2	<input type="text"/>	
Руб. на 1 ПГИ	<input type="text"/>	Сценарий 3	<input type="text"/>	
Дней на 1 ПГИ	<input type="text"/>	Сценарий 4	<input type="text"/>	
Руб. на 1 РИР	<input type="text"/>	Сценарий 5	<input type="text"/>	
Дней на 1 РИР	<input type="text"/>	Сценарий 6	<input type="text"/>	

Критические уровни падения дебита нефти и роста дебита воды для промежутка в 2 года

Δ нефти  Δ воды

**Заполнены не все обязательные поля**

Рис.3.6 Интерфейс расчетной программы для кандидатов РИР.

(материальные и временные затраты, вероятности сценариев ремонтных работ, критические величины изменения дебита). Интерактивное окно содержит инструкции для пользователя, касающиеся последовательности действий при работе с программой и формата вводимых данных.

В скрипке осуществляется проверка полноты и формата вводимых данных. При обнаружении незаполненных обязательных полей, текстовых значений вместо числовых или некорректной суммы вероятностей для сценариев ремонта программа выводит соответствующее предупреждение для пользователя в интерактивном окне и не позволяет запустить основной скрипт. В консоль выводится информация о всех вызываемых функциях модуля и

значения переменных, вводимых пользователем.

### **Краткое описание основного скрипта.**

В основном скрипте полученные из пользовательского ввода значения импортируются, присваиваются соответствующим переменным и используются в процессе извлечения и обработки данных.

В консоль выводится информация о всех вызываемых функциях основного скрипта.

В корневой папке осуществляется перебор всех вложенных элементов с составлением списка путей к релевантным файлам: документам MS Excel, названия которых содержат ключевое слово “Шахматка”.

Из каждого найденного файла извлекается несколько наборов данных:

1. Ежедневные данные по параметрам “Qн”, “Qж”, “Обв” и “Примечание” (показан преобразованный формат):

Параметр	Qж	Qн	Обв	Примечание
Дата				
2019-12-12	18.0	3.29	77.3	0.0
2019-12-13	17.0	3.11	77.3	0.0
2019-12-14	18.0	3.29	77.3	0.0
2019-12-15	17.0	3.11	77.3	0.0
2019-12-16	17.0	3.11	77.3	0.0

Таблица 2. Вид вводимых данных для расчетной программы кандидатов для РИР.

Примечания исходного файла преобразуются в бинарный числовой столбец, где 1 - наличие любого примечания, 0 - отсутствие примечаний. Данные могут быть использованы для создания моделей МО.

2. Среднемесячные данные по всем заполненным параметрам (показан исходный формат):



	A	B	C	D	E	F	G	H	I	
1	Куст	341	5150				00 Забой иск. 3958Ал			
2	ЗУ									
3			Цех		ЦДНГ-7					
4	Насос	Дата	Режимы		Параметр	Средн.	1	2	3	
467	00, глуб. 3050 м., МРП 297 Д.,	Февраль 2018			Qн ТМ	58,48	37,93	41,7	40,7	
468					Ф вращ ТМ	49,18	53,5	53,5	53,5	
469					Qж(тн)	481,19	489,6	484,6	484,6	
470					Qн	9,59	9,55	9,45	9,45	
471					Обв	98,13				
472					Qж	481,94	487	482	482	
473					Обв ХАЛ	97,95				
474					Qж ТМ	477,13	487	484	482	
475					Ф	54	54			
476					Дшт	32	32			
477					Qж ТМ (исх)	476,06	487	484	482	
478					Рбуф ТМ					
479			Qг (АГЗУ)							

Рис.3.7 Вид исходного формата «шахматок».

Данные используются для статистической оценки NPV в состоянии “как есть” и с учетом возможного проведения ремонтных работ.

3. Данные из шапки экселевских таблиц, содержащие показатель накопленной добычи нефти, дату последней перфорации и ГРП (показан исходный формат):

L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF
накоп. добыча нефти, тн			интервал перфорации, дни			Перфорация, м			Дата посл. перф.			Удл	Газ. фактор	Рпл	НКТ		ГРП			
11563,5540			4			3,99;3723,32 - 3 муфты;перф			10.03.2017			1098	282	14,73323	№ 3 6		03.2017 (ОАО "ИНКОМнефть)			
6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
41,1	40,97	43,67	40,87	42,03	42	43,35	40,67	42,3	41,6	43,7	40	38,48	39,43	40,65	200,2	50,93	53,3	178,6	59,75	140,1
53,5	53,5	53,5	53,5	53,5	53,5	53,5	53,5	53,5	53,5	53,5	53,5	53,5	53,5	53,5	54	54	54	54	54	54
484,9	300,7	486,1	484,1	487,1	487,8	487,8	486,5	489,5	489,5	489,5	489,5	487,5	487,5	489,5	469,4	498,6	502,3	505	501	470,7
8,27	9,8	15,85	15,78	15,88	13,1	13,1	9,88	9,95	9,95	9,95	9,95	9,9	9,9	9,95	9,54	10,13	11,44	0	0	0
	96				96,7		97,5		97,5								97,2	100	100	100
482	300	485	483	486	486	486	484	487	487	487	487	485	485	487	467	496	500	500	496	466
					96,7		97,5		98,4								97,2			
482	482	485	485	486	486	486	484	487	487	487	487	485	485	454	467	496	500	500	498	466
482	482	485	485	486	486	486	484	487	487	487	487	485	485	454	382	496	500	435	498	466

Рис.3.8 Входные данные из «шахматок».

В процессе обработки данных формируется таблица (датафрейм), содержащая укрупненные показатели для всех рассматриваемых скважин. Таблица содержит следующие поля:

- 'Q\_oil' - средний дебит нефти за последние 6 мес., т в сутки;
- 'Max\_oil' - максимальный дебит нефти в предшествующий период (в пределах 2-х последних лет), т в сутки в среднем за 1 месяц;
- 'Oil\_delta' - изменение дебита нефти в долях от 1 ( $Q_{oil} / Max_{oil}$ );
- 'Oil\_total' - накопленная добыча нефти, т;

- 'Q\_water' - средний дебит воды за последние 6 мес., т в сутки;
- 'Min\_water' - минимальный дебит воды в предшествующий период (в пределах 2-х последних лет), т в сутки в среднем за 1 месяц;
- 'Water\_delta' - изменение дебита воды в долях от 1 ( $Q\_water / Min\_water$ );
- 'Water\_pc' - средняя обводненность за последние 6 мес., %;
- 'Min\_water\_pc' - минимальный процент обводненности в предшествующий период (в пределах 2 последних лет);
- 'Water\_pc\_delta' - изменение обводненности в долях от 1 ( $Water\_pc / Min\_water\_pc$ );
- 'NPV\_as\_is' - ожидаемый доход (убыток) при отсутствии ремонтных работ;
- 'NPV\_repair' - ожидаемый доход (убыток) от проведения ремонтных работ;
- 'Scenario\_1' <...> 'Scenario\_6' - ожидаемый доход (убыток) по сценариям;
- 'Last\_perf' - дата последней перфорации;
- 'GRP' - дата ГРП;
- 'Defect' - наличие повреждения скважины (True - вероятно, False - не обнаружено);
- 'Repair' - рекомендации по ремонту (True - ремонт рекомендуется, False - не рекомендуется);
- 'Path' - путь к файлу с шахматкой скважины.

Все собранные данные после завершения обработки сохраняются в файл "Candidates.xlsx". Файл автоматически открывается после создания.

### Алгоритм статистической оценки NPV в состоянии “как есть”.

1. Вычисляется среднесуточное значение дебита нефти (“Q\_oil”) и воды (“Q\_water”) за последние 6 месяцев:

$$Q_{cp} = \sum \frac{Q_n}{n} \quad (3.1)$$

где n = 6, количество месяцев

2. Рассчитывается продолжительность прогнозного периода работы скважины в днях:

$$prod\_days = 6 / 12 * 365 \quad (3.2)$$

3. Рассчитывается показатель NPV без проведения ремонтных работ:

$$NPV\_as\_is = Q\_oil * prod\_days * oil\_price - Q\_water * prod\_days * water\_cost \quad (3.3)$$

где:

Q\_oil - среднесуточный дебит нефти за последние 6 месяцев;

Q\_water - среднесуточный дебит воды за последние 6 месяцев;

prod\_days - количество продуктивных дней работы скважины;

oil\_price - денежные поступления от 1 т добытой нефти;

water\_cost - денежные затраты на утилизацию 1 т воды.

### Алгоритм идентификации возможного дефекта скважины.

1. Рассчитывается отношение среднесуточного дебита нефти и воды за последние 6 месяцев к соответствующим максимальным и минимальным среднемесячным показателям за предшествующие 1,5 года.

Таблица 3. Параметры для расчета кандидатов РИР.

Последние 2 года		Коэффициент (отношение)
Предшествующие 1,5 года	Последние 6 мес.	
Максимальный среднемесячный дебит нефти, т/сутки (Max_oil)	Средний дебит нефти, т/сутки (Q_oil)	$Q\_oil / Max\_oil$
Минимальный среднемесячный дебит воды, т/сутки (Min_water)	Средний дебит воды, т/сутки (Q_water)	$Q\_water / Min\_water$

2. Полученные коэффициенты (отношения текущих показателей к пиковым значениям на временном интервале в 2 года) сравниваются с критическими показателями, указанными пользователем в интерактивном окне при запуске программы.

Если падение дебита нефти или рост дебита воды превышает критические значения, делается вывод о том, что в скважине вероятно наличие дефекта и требуется оценка эффективности ремонтных работ.

Если изменение дебита нефти и воды не превышает критических значений, делается вывод о том, что ремонт скважины не требуется. Оценка NPV для сценариев ремонтных работ на таких скважинах не проводится.

Соответствующие значения (True или False) вносятся в столбец “Defect” сводной таблицы.

### **Алгоритм статистической оценки NPV при проведении ремонтных работ.**

1. Из общего перечня скважин отбираются скважины, имеющие значение True в столбце “Defect” сводной таблицы (скважины с вероятными дефектами).
2. Для каждой скважины осуществляется расчет показателя NPV по 6-ти сценариям ремонтных работ:
  - Сценарий 1: 1 ПГИ, добыча не изменилась;
  - Сценарий 2: 2 ПГИ, добыча не изменилась;
  - Сценарий 3: 2 ПГИ, 1 РИР, добыча нефти выросла;
  - Сценарий 4: 3 ПГИ, 1 РИР, добыча нефти выросла;
  - Сценарий 5: 3 ПГИ, 2 РИР, добыча нефти выросла;
  - Сценарий 6: 4 ПГИ, 2 РИР, добыча нефти выросла.

Для каждого сценария последовательно рассчитывается продолжительность прогнозного периода работы скважины в днях (“prod\_days”) с учетом временных затрат на проведение каждого РИР и ПГИ.

Расчет NPV по каждому сценарию предусматривает материальные затраты на соответствующее количество РИР и ПГИ.

Первые два сценария не предполагают нормализации дебита после ремонта скважины, сохраняется средний за последние 6 месяцев уровень дебита.

Сценарии 3-6 предусматривают, что после ремонта дебит нефти увеличится до уровня, соответствующего 95% от предшествующего пикового значения, дебит воды сократится до 105% от предшествующего минимального показателя.

Поправка в 5% к предшествующим оптимальным значениям дебита введена

для того, чтобы отразить процессы естественных (неустраняемых) изменений в скважине.

$$\text{Scenario}_1' = Q_{\text{oil}} * \text{prod\_days} * \text{oil\_price} - Q_{\text{water}} * \text{prod\_days} * \text{water\_cost} - \text{PGI\_cost} \quad (3.4)$$

$$\text{Scenario}_2 = Q_{\text{oil}} * \text{prod\_days} * \text{oil\_price} - Q_{\text{water}} * \text{prod\_days} * \text{water\_cost} - \text{PGI\_cost} * 2 \quad (3.5)$$

$$\text{Scenario}_n = \text{Max\_oil} * 0.95 * \text{prod\_days} * \text{oil\_price} - \text{Min\_water} * 1.05 * \text{prod\_days} * \text{water\_cost} - \text{PGI\_cost} * \text{pn} - \text{RIR\_cost} * \text{rn} \quad (3.6)$$

где:

$Q_{\text{oil}}$  - среднесуточный дебит нефти за последние 6 месяцев;

$Q_{\text{water}}$  - среднесуточный дебит воды за последние 6 месяцев;

$\text{prod\_days}$  - количество продуктивных дней работы скважины;

$\text{oil\_price}$  - денежные поступления от 1 т добытой нефти;

$\text{water\_cost}$  - денежные затраты на утилизацию 1 т воды;

$\text{PGI\_cost}$  - материальные затраты на 1 ПГИ;

$\text{pn}$  - количество проводимых ПГИ;

$\text{RIR\_cost}$  - материальные затраты на 1 РИР;

$\text{rn}$  - количество проводимых РИР.

3. На основе полученных оценок NPV по 6-ти сценариям ремонтных работ и вероятности каждого сценария рассчитывается ожидаемый показатель NPV\_repair:

$$NPV_{\text{repair}} = \sum(\text{Scenario}_n * P_n) \quad (3.7)$$

Где Scenario\_n - ожидаемый показатель NPV по сценариям;

Pn - вероятности сценариев в долях от 1.

### **Алгоритм итоговых рекомендаций по проведению ремонтных работ.**

1. На основе полученных оценок NPV рекомендуется ремонт скважин, у которых ожидаемый показатель NPV с учетом вероятностей исхода ремонтных работ является положительным и превышает NPV в текущем состоянии (без ремонта):

$$NPV_{\text{repair}} > 0 \quad (3.8)$$

$$NPV_{\text{repair}} > NPV_{\text{as\_is}} \quad (3.9)$$

Соответствующим образом заполняется столбец “Repair”: значение True присваивается скважинам, ремонт которых целесообразен, False - скважинам, ремонт которых не рекомендуется.

2. Сводный набор данных сортируется по значениям столбцов “Repair” и “NPV\_repair” по убыванию значений. Таким образом в итоговой таблице, сохраняемой в файл, в верхней части идут скважины, рекомендуемые к ремонту и имеющие наиболее высокий ожидаемый NPV при проведении ремонтных работ. В конце таблицы идут скважины с минимальными и отрицательными значениями “NPV\_repair” и скважины без дефектов, ремонт которых не требуется.

## ЗАКЛЮЧЕНИЕ

В результате данного исследования был проведен анализ различных методик VOI с приложениями к нефтяной промышленности. Был определен спектр задач, к которым можно применить данную методологию и была определена новая задача применимости методологии VOI для оценки экономической целесообразности проведения промысловых геофизических исследований для ремонтно-изоляционных работ на основе информации исторических данных о скважине. В процессе данного исследования были реализованы следующие практические результаты:

- Был проведен анализ входных данных с добывающих скважин и разработан алгоритм анализа подготовки данных для последующего использования их в исследовании.
- Были разработаны сценарии работы скважин для дальнейшего использования этой информации для расчета суммы дисконтированных денежных потоков платежей к сегодняшнему дню, с учетом и без учета дополнительной информации о скважине, для дальнейшего расчета параметра ценности информации.
- Были разработаны алгоритмы расчета ожидаемых ценностей проекта с учетом и без учета дополнительной информации о скважинах и визуализированы в виде графической расчетной блок-схемы.
- Был проведен анализ работ алгоритмов машинного обучения, были подобраны оптимальные параметры работы этих алгоритмов и были выбраны наиболее подходящие из этих алгоритмов машинного обучения для дальнейших расчетов ожидаемых ценностей проекта с учетом и без учета дополнительной информации.
- Реализован программный комплекс для оценки экономической целесообразности проведения промысловых геофизических



исследований для ремонтно-изоляционных работ на основе исторических данных о скважине.

Это исследование только начальная стадия разработки методологии VOI для задачи оценки экономической целесообразности проведения ПГИ для РИР на основе экономических данных о скважине. В данном исследовании на конечный результат учитывалось только влияние предсказанных дебитов нефти и дебитов воды и не учитывались другие факторы. Хотя на конечный результат может оказывать влияние нагнетательные скважины, а также геометрия расположений нагнетательных скважин, и, кроме того, другие параметры могут оказывать влияние на конечный расчет VOI. Чтобы составить более качественную методологию VOI для исследуемой задачи, необходимо учитывать большое количество неопределенностей, что требует более масштабного исследования поставленной задачи.

## СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. Abramowitz M., I.A. Stegun, Handbook of Mathematical Functions. New York: Dover Publications Inc., 1972.
2. Aggrey, G.H., Davies, D.R., Ajayi, A., and Konopczynski, M. 2006. Data Richness and Reliability in Smart-Field Management—Is There Value? Paper SPE 102867 presented at the SPE Annual Technical Conference and Exhibition, San Antonio, Texas, 24–27 September.
3. Ballin, P.R., Ward, G.S., Whorlow, C.V., and Kahn, T. 2005. Value of Information for a 4D-Seismic Acquisition Project. Paper SPE 94918 presented at the SPE Latin American and Caribbean Petroleum Engineering Conference, Rio de Janeiro, Brazil, 20–23 June.
4. Bickel, J.E. and Bratvold, R.B. 2007. Decision Making in the Oil & Gas Industry: From Blissful Ignorance to Uncertainty-Induced Confusion Paper SPE 109610 presented at the SPE Annual Technical Conference and Exhibition, Anaheim, California, 11–14 November.
5. Bickel, J.E., Gibson, R.L., McVay, D.A., Pickering, S., and Waggoner, J. 2006. Quantifying 3D Land Seismic Reliability and Value. Paper SPE 102340 presented at the SPE Annual Technical Conference and Exhibition, San Antonio, Texas, 24–27 September.
6. Blackwell, D. 1953. Equivalent Comparisons of Experiments. *The Annals of Mathematical Statistics* 24(2), 265- 272.
7. Branco, C.C.M., Pinto, A.C.C., Tinoco, P., M.B., Vieira,
8. Bratvold, R.B. and Begg, S.H. 2006. Education for the Real World: Equipping Petroleum Engineers to Manage Uncertainty. Paper SPE 103339 presented at the SPE Annual Technical Conference and Exhibition, San Antonio, Texas, 24–27 September.
9. *British Journal for the Philosophy of Science* 41 1-4.

10. Churchill R.V. Operational Mathematics. New York: McGraw-Hill Book Co. Inc., 1972.
11. Churchill R.V. Operational Mathematics. New York: McGraw-Hill Book Co. Inc., 1972.
12. Clemen, R.T. and Reilly, T. 2001. Making Hard Decisions.
13. Clemen, R.T. and Winkler, R.L. 1985. Limits for the Precision and Value of Information from Dependent Sources. Operations Research 33 (2) 427-442.
14. Coopersmith, E.M. and Cunningham, P.C. 2002. A Practical Approach to Evaluating the Value of Information and Real Option Decisions in the Upstream Petroleum Industry. Paper SPE 77582 presented at the SPE Annual Technical Conference and Exhibition, San Antonio, Texas, 29 September–2 October.
15. Coopersmith, E.M., Burkholder, M.K., and Schulze, J.H. 2006. Value of Information Lookbacks—Was the Information You Gathered Really Worth Getting? Paper SPE 101540 presented at the SPE Annual Technical Conference and Exhibition, San Antonio, Texas, 24-27 September.
16. Demirmen, F. 1996. Use of "Value of Information" Concept in Justification and Ranking of Subsurface Appraisal.
17. Demirmen, F. 2001. Subsurface Appraisal: The Road from Reservoir Uncertainty to Better Economics. Paper SPE 68603 presented at the SPE Hydrocarbon Economics and Evaluation Symposium, Dallas, Texas, 2–3 April.
18. Dougherty, E.L., 1971. The Oilman's Primer on Statistical Decision Theory, in SPE Library (unpublished).
19. Dunn, M.D. 1992. A Method to Estimate the Value of Well Log Information. Paper SPE 24672 presented at the SPE Annual Technical Conference and Exhibition, Washington, DC, 4–7 October.
20. Galli, A., Armstrong, M., and Dias, M.A.G. 2004. The Value of Information: A Bayesian Real Option Approach. Paper SPE 90418

- presented at the SPE Annual Technical Conference and Exhibition, Houston, Texas, 26–29 September.
21. Gerhardt, J.H. and Haldorsen, H.H. 1989. On the value of information. Paper SPE 19291 presented at the 1989 Offshore Europe, Aberdeen, 5-8 September.
  22. Gould, J.P. 1974. Risk, Stochastic Preference, and the Value of Information. *Journal of Economic Theory* 8 (64) 64-84.
  23. Grayson C. J. *Decision Under Uncertainty: Drilling Decisions by Oil and Gas Operators*. 1960.
  24. Grayson, C.J., Jr. 1962. Bayesian Analysis—A New Approach to Statistical Decision-Making. *Journal of Petroleum Technology* (June) 603-607.
  25. Gringarten A.C. and Ramey. H.J. *The Use of Source and Green's Functions in Solving Unsteady-Flow Problems in Reservoirs*. b.m.: SPEJ, 1973.
  26. Haskett, W.J. 2003. Optimal Appraisal Well Location Through Efficient Uncertainty Reduction and Value of Information Techniques. Paper SPE 84241 presented at the SPE Annual Technical Conference and Exhibition, Denver, Colorado, 5–8 October.
  27. Head, K.J. 1999. How Could You Possibly Predict the Value of 3-D Seismic Before You Shoot It? Paper SPE 56446 presented at the SPE Geo-Triad Convention, Calgary, Canada, 15–19 June.
  28. Hilton, R.W. 1979. The determinants of cost information value: An illustrative analysis. *Journal of Accounting Research* 17 (2) 411-435.
  29. Hilton, R.W. 1981. The Determinants of Information Value: Synthesizing Some General Results. *Management Science* 27 (1) 57-64.
  30. Hirakawa, S. 1972. Risk Analysis in the Petroleum Production Industry. Paper SPE 4295 presented at the SPE Joint AIME-MMIJ Meeting, Tokyo, Japan, 25–27 May.
  31. Hirschleifer, J. 1961. The Bayesian Approach to Statistical Decision: An Exposition. *Journal of Business* 34 (4) 471-489.
  32. Houzé O., Viturat D., Fjaere O-S., *Dynamic data analysis*. b.m.: Kappa, 2008.

- 33.Howard, R.A. 1966a. Decision Analysis: Applied Decision Theory. In Proceedings of the Fourth International Conference on Operational Research, ed. D.B. Hertz and J. Melese, 55-71. New York, NY. Wiley-Interscience.
- 34.Howard, R.A. 1966b. Information Value Theory. IEEE Transactions on Systems Science and Cybernetics SSC-2 (1) 22-26.
- 35.Howard, R.H. 2005. Decision Analysis Manuscript: unpublished.
- 36.Jacobs, J. and Ward, R. 2006. Digital Oil Field of the Future. Paper SPE presented at the CERAWeek, Houston, Texas, 6–10 February.
- 37.Jan R. Williams, Susan F. Haka, Mark S. Bettner, Joseph V. Carcello.,. Financial & Managerial Accounting. 2008.
- 38.Joshi S.D. Horizontal Well Technolog. Tulsa, Oklahoma: Penwell Published Compan, 1991.
- 39.Kihlstrom, R. 1974. A general theory of demand for information about product quality. Journal of Economic Theory 8 (4) 413-439.
- 40.Koninx, J-P.M. 2001. Value-of-Information: from Cost- cutting to Value-Creation. Journal of Petroleum Technology 53 (4) 84-92.
- 41.Kumar, R., and Hara, S.K. 2005. Value of Information for Appraisal of Multiple Dependent Prospects. Paper SPE 96171 presented at the SPE Annual Technical Conference and Exhibition, Dallas, Texas, 9–12 October.
- 42.Lohrenz, J. 1988. Net Values of Our Information. Journal of Petroleum Technology 40 (4) 499-503.
- 43.Moras, R.G., Lesso, W.G., and MacDonald, R.C. 1987. Assessing the Value of Information Provided by Observation Wells in Gas Storage Reservoirs, in SPE Library (unpublished).
- 44.P.M.F., Sayd, A.M., Santos, R.L.A., and Prais, F. 2005. The Role of the Value of Information and Long Horizontal Wells in the Appraisal and Development Studies of a Brazilian Offshore Heavy-Oil Reservoir Paper SPE 97846 presented at the SPE International Thermal Operations and Heavy Oil

- Symposium, Calgary, Alberta, Canada, 1-3 November.
45. Pacific Grove, CA: Duxbury.
  46. Paper SPE 36631 presented at the 1996 Annual Technical Conference and Exhibition, Denver, Colorado, 6–9 October.
  47. Portella, R.C.M., Salomao, M.C., Blauth, M., and Duarte,
  48. Prange, M., Armstrong, M., Bailey, W., Couet, B., Djikpesse, H., Wilkinson, D., and Galli, A. 2006. Better Valuation of Future Information Under Uncertainty. Paper SPE 103028 presented at the SPE Annual Technical Conference and Exhibition, San Antonio, Texas, 24–27 September.
  49. R.L.B. 2003. Uncertainty Quantification to Evaluate the Value of Information in a Deepwater Reservoir. Paper SPE 79707 presented at the SPE Reservoir Simulation Symposium, Houston, Texas, 3–5 February.
  50. Raghuraman, B., Couet, B., and Savundararaj, P. 2002. Valuation of Technology and Information for Reservoir Risk Management. Paper SPE 77424 presented at the SPE Annual Technical Conference and Exhibition, San Antonio, Texas, 29 September–2 October 2002.
  51. Raiffa, H. 1968. *Decision Analysis: Introductory Lectures on Choices under Uncertainty*. Reading, MA: Addison- Wesley.
  52. Raiffa, H. and Schlaifer, R. 1961. *Applied Statistical Decision Theory*. Cambridge, MA: Harvard University.
  53. Ramsey, F.P. 1926. Truth and Probability. In *Philosophical Papers*, ed. D.H. Mellor, 52-94, Cambridge, England. Cambridge University Press.
  54. Ramsey, F.P. 1990. *Weight or the Value of Knowledge*.
  55. *Reservoir Evaluation and Engineering* 3 (4) 310-318. Waggoner, J.R. 2002. Quantifying the Economic Impact of 4D
  56. Samson, D., Wirth, A., and Rickard, J. 1989. The value of information from multiple sources of uncertainty in decision analysis. *European Journal of Operational Research* 39 254-260.
  57. Savage, L.J. 1954. *The Foundations of Statistics*. New York, NY: John Wiley & Sons.

58. Schlaifer, R. 1959. *Probability and Statistics for Business Decisions*. New York, NY: McGraw-Hill Book Company, Inc.
59. Schlaifer, R. 1961. *Introduction to Statistics for Business Decisions*. New York, NY: McGraw-Hill.
60. Seismic Projects. *SPE Reservoir Evaluation and Engineering* (April) 111-115.
61. Shannon, C.E. 1948. A Mathematical Theory of Communication. *The Bell System Technical Journal* 37 (3) 379-423, 623-656.
62. Silbergh, M. and Brons, F. 1972. Profitability Analysis - Where Are We Now? *Journal of Petroleum Technology* 24 (1) 90-100.
63. Steagall, D.S., Gomes, J.A.T., Oliveira, R.M., Ribeiro, N.M.S.J., Queiroz, R.Q., Carvalho, M.R.J., and Souza, C.Z. 2005. How to Estimate the Value of Information (VOI) of a 4D Seismic Survey in One Offshore Giant Field. Paper SPE 95876 presented at the SPE Annual Technical Conference and Exhibition, Dallas, Texas, 9–12 October.
64. Stehfest H. Numerical Inversion of Laplace Transforms. b.m.: Communications of the ACM, 1970.
65. Stibolt, R.D. and Lehman, J. 1993. The Value of a Seismic Option. Paper SPE 25821 presented at the 1993 SPE Hydrocarbon Economics and Evaluation Symposium, Dallas, Texas, 29–30 March.
66. Van Everdingen A.F. and Hurst W. W. The Application of the Laplace Transformation to Flow Problems in Reservoirs. b.m.: AIME, 1953.
67. Waggoner, J.R. 2000. Lessons Learned From 4D Projects.
68. Warren. 1983. Development Decision: Value of Information. Paper SPE 11312 presented at the SPE Hydrocarbon Economics and Evaluation Symposium, Dallas, Texas, 3–4 March.
69. Wills, H.A. and Graves, R.M. 2004. Information is Costly, but How Valuable is It? Paper SPE 90710 presented at the SPE Annual Technical Conference and Exhibition, Houston, Texas, 26–29 September.

70. Wilson, R. 1975. Informational Economies of Scale. *Bell Journal of Economics* 6 (1) 184-195.
71. Winkler, R.L. 2003. *An Introduction to Bayesian Inference and Decision*. Gainesville, FL: Probabilistic Publishing.
72. Байков В.А., Бакиров Н. К., Яковлев А. А., *Математическая геология*. Москва-Ижевск: б. н., 2012. стр. 228. Т. 1: Введение в геостатистику.
73. *Геология нефти и газа: Учебник для вузов* / Э. А. Бакиров, В. И. Ермолкин, В. И. Ларин и др. // Под ред. Э. А. Бакирова. – М.: Недра, 1990. – 240 с.
74. Демьянов В. В., Савельева Е. А. *Геостатистика: теория и практика*. [ред.] Р. В. Арутюняна. Москва, 2010. стр. 327.
75. Каневский М.Ф., Демьянов В. В., Савельева Е. А. и др., *Элементарное введение в геостатистику*. Москва: ВИНТИ, 1999.
76. Кременецкий М. И., Ипатов А. И., Гуляев Д. Н. *Информационное обеспечение и технологии гидродинамического моделирования нефтяных и газовых залежей*. Москва-Ижевск: Ижевский институт компьютерных исследований, 2011. стр. 896.
77. Ортега Дж. *Введение в параллельные и векторные методы решения линейных систем*. Москва: Мир, 1991.



## ПРИЛОЖЕНИЕ 1

### Обозначения и сокращения.

- *ПГИ* – промыслово-геофизические исследования;
- *ГТМ* – геолого-технологические мероприятия
- *РИР* – ремонтно-изоляционные работы
- *ВНК* – водонефтяной коллектор
- *ГНК* – газонефтяной коллектор
- *ОПЗ* – область призабойной зоны
  
- $Q_n(t)'$  – прогноз накопленного за время  $t$  дебита нефти при текущем состоянии скважины.
- $Q_g(t)'$  – прогноз накопленной за время  $t$  дебита воды при текущем состоянии скважины.
  
- $C_n$  – предполагаемый доход от 1 тонны нефти.
- $C_g$  – предполагаемый расход на транспортировку 1 тонны воды.
  
- $Q_n(t)''$  – прогноз накопленного за время  $t$  дебита нефти после проведения РИР.
- $Q_g(t)''$  – прогноз накопленного за время  $t$  дебита воды после проведения РИР.
  
- $Q_n(t_{nzu})'$  – прогноз накопленного дебита нефти на момент проведения ПГИ.
- $Q_n(t_{nzu2})'$  – прогноз накопленного дебита нефти на момент повторного проведения ПГИ.

- $Q_n(t_{ngu}')'$  – прогноз накопленного дебита нефти после РИР на момент проведения повторного ПГИ.
- $Q_n(t_{ngu}'')$  – прогноз накопленного дебита нефти на момент второго проведения ПГИ.
- $Q_n(t_{rip})'$  – прогноз накопленного дебита нефти на момент проведения РИР.
- $Q_n(t_{rip2})'$  – прогноз накопленного дебита нефти на момент проведения повторного РИР.
- $C_{ngu}$  – предполагаемый расход на проведение ПГИ.
- $C_{ngu2}$  – предполагаемый расход на проведение повторного ПГИ.
- $C_{ngu}'$  – предполагаемый расход на проведения ПГИ после первичного РИР.
- $C_{ngu}''$  – предполагаемый расход на проведение ПГИ после вторичного РИР.
- $C_{rip}$  – предполагаемые расходы на проведение РИР.
- $C_{rip2}$  – предполагаемые расходы на проведения повторного РИР.
- $Q_n$  – посуточная история дебита нефти.
- $Q_{ж}$  – посуточная история дебита жидкости.
- $t_{ngu}$  – предполагаемая продолжительность проведения ПГИ.
- $t_{ngu2}$  – предполагаемая продолжительность проведения повторного ПГИ.

- $t_{rip}$  – предполагаемая продолжительность проведения РИР.
- $t_{rip2}$  – предполагаемая продолжительность проведения повторного РИР.
- $t$  – продолжительность дебита в месяцах по которой планируется определять VOI.
- $P_i$  – вероятность происхождения  $i$ -го исхода.
- НД – набор данных
- ГС – горизонтальные скважины
- ПЭВМ – персональная электронная вычислительная машина