TOMSK POLYTECHNIC UNIVERSITY ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский Томский политехнический университет» (ТПУ)

Школа Информационных технологий и робототехники
Направление подготовки 09.04.04 Программная инженерия
Отделение школы (НОЦ) Информационных технологий

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

Тема работы
Методология подготовки исходных данных для модели машинного обучения в нефтегазовой области

УДК 004.853:622.32

Студент

Группа	ФИО	Подпись	Дата
8ПМ8И	Журбич Никита Игоревич		10.06.2020

Руководитель ВКР

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОИТ ИШИТР	Губин Евгений Иванович	к.ф.-м.н.		10.06.2020

КОНСУЛЬТАНТЫ ПО РАЗДЕЛАМ:

По разделу «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение»

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОСГН ШБИП	Меньшикова Екатерина Валентиновна	к.ф.н.		02.06.2020

По разделу «Социальная ответственность»

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ООД ШБИП	Горбенко Михаил Владимирович	к.т.н		10.06.2020

ДОПУСТИТЬ К ЗАЩИТЕ:

Руководитель ООП	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОИТ ИШИТР	Губин Евгений Иванович	к.ф.-м.н.		10.06.2020

Томск – 2020 г.

TOMSK POLYTECHNIC UNIVERSITY ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский Томский политехнический университет» (ТПУ)

Школа Информационных технологий и робототехники
Направление подготовки (специальность) 09.04.04. Программная инженерия
Отделение школы (НОЦ) Информационных технологий

УТВЕРЖДАЮ:
Руководитель ООП
_____ 10.05.2020 Губин Е.И.
(Подпись) (Дата) (Ф.И.О.)

ЗАДАНИЕ
на выполнение выпускной квалификационной работы

В форме:

Магистерской диссертации

(бакалаврской работы, дипломного проекта/работы, магистерской диссертации)

Студенту:

Группа	ФИО
8ПМ8И	Журбич Никита Игоревич

Тема работы:

Методология подготовки исходных данных для модели машинного обучения в нефтегазовой области	
Утверждена приказом директора (дата, номер)	59-62/С

Срок сдачи студентом выполненной работы:	13.06.2020

ТЕХНИЧЕСКОЕ ЗАДАНИЕ:

Исходные данные к работе	Объектом исследования является методология обработки исходных данных с нефтяного месторождения, с помощью которой можно построить прогнозные модели для дальнейших исследований. Среды разработки программного обеспечения Jupyter Notebook и R Studio. Языки программирования Python и R.

Перечень подлежащих исследованию, проектированию и разработке вопросов	1. Обзор предметной области 2. Проектирование алгоритмов программного комплекса. 3. Разработка алгоритмов в виде программного комплекса 4. Финансовый менеджмент, ресурсоэффективность и ресурсосбережение 5. Социальная ответственность
Перечень графического материала	1. Диаграмма Исикавы 2. Схема приоритетов внедрения технологий Big Data в нефтегазовом секторе 3. Матрица корреляций

Консультанты по разделам выпускной квалификационной работы
(с указанием разделов)

Раздел	Консультант
Финансовый менеджмент, ресурсоэффективность и ресурсосбережение	Меньшикова Екатерина Валентиновна
Социальная ответственность	Горбенко Михаил Владимирович
Раздел на английском языке	Пичугова Инна Леонидовна

Названия разделов, которые должны быть написаны на русском и иностранном языках:

1. Обзор предметной области

Дата выдачи задания на выполнение выпускной квалификационной работы по линейному графику	01.03.2020

Задание выдал руководитель:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОИТ ИШИТР	Губин Евгений Иванович	к.ф.-м.н.		10.06.2020

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
8ПМ8И	Журбич Никита Игоревич		10.06.2020

TOMSK POLYTECHNIC UNIVERSITY ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский Томский политехнический университет» (ТПУ)

Школа Информационных технологий и робототехники
Направление подготовки (специальность) 09.04.04 Программная инженерия
Уровень образования Магистратура
Отделение школы (НОЦ) Информационных технологий
Период выполнения Весенний семестр 2019 /2020 учебного года

Форма представления работы:

Магистерская диссертация

(бакалаврская работа, дипломный проект/работа, магистерская диссертация)

КАЛЕНДАРНЫЙ РЕЙТИНГ-ПЛАН
выполнения выпускной квалификационной работы

Срок сдачи студентом выполненной работы:	10.06.2020

Дата контроля	Название раздела (модуля) / вид работы (исследования)	Максимальный балл раздела (модуля)
10.03.2020	Раздел 1. Обзор предметной области	20
10.04.2020	Раздел 2. Разведочный анализ данных	20
15.05.2020	Раздел 3. Регрессионный анализ	30
30.05.2020	Раздел 4. Финансовый менеджмент, ресурсоэффективность и ресурсосбережение	15
30.05.2020	Раздел 5. Социальная ответственность	15

СОСТАВИЛ:
Руководитель ВКР

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОИТ ИШИТР	Губин Евгений Иванович	к.ф.-м.н.		10.06.2020

СОГЛАСОВАНО:

Руководитель ООП

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОИТ ИШИТР	Губин Евгений Иванович	к.ф.-м.н.		10.06.2020

ЗАДАНИЕ ДЛЯ РАЗДЕЛА
«ФИНАНСОВЫЙ МЕНЕДЖМЕНТ, РЕСУРСОЭФФЕКТИВНОСТЬ И РЕСУРСОСБЕРЕЖЕНИЕ»

Студенту:

Группа	ФИО
8ПМ8И	Журбич Никита Игоревич

Школа	ИШИТР	Отделение школы (НОЦ)	ОИТ
Уровень образования	Магистр	Направление/специальность	09.04.04 «программная инженерия»

Исходные данные к разделу «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение»:	
1. *Стоимость ресурсов научного исследования (НИ): материально-технических, энергетических, финансовых, информационных и человеческих*	Стоимость материальных ресурсов определялась согласно прейскурантам компаний Оклад руководителя – 33664 р. Оклад инженера – 21760 р.
2. *Нормы и нормативы расходования ресурсов*	Накладные расходы 16 %; Районный коэффициент 30%; Норма амортизации ПЭВМ 33,33%; Норма амортизации ПО 20%
3. *Используемая система налогообложения, ставки налогов, отчислений, дисконтирования и кредитования*	Коэффициент отчислений на уплату во внебюджетные фонды 30%
Перечень вопросов, подлежащих исследованию, проектированию и разработке:	
1. *Оценка коммерческого и инновационного потенциала НТИ*	Анализ потенциальных потребителей результатов исследования, оценка качества и перспективности проекта по технологии QuaD, SWOT-анализ
2. *Разработка устава научно-технического проекта*	Инициация проекта: определение заинтересованных сторон проекта, целей и результатов проекта
3. *Планирование процесса управления НТИ: структура и график проведения, бюджет, риски и организация закупок*	План проекта, определение трудоемкости выполнения работ, разработка графика проведения научного исследования, расчет бюджета разработки
4. *Определение ресурсной, финансовой, экономической эффективности*	Описание потенциального эффекта
Перечень графического материала *(с точным указанием обязательных чертежей):*	
1. Матрица SWOT 2. График проведения и бюджет НТИ 3. Оценка ресурсной, финансовой и экономической эффективности НТИ 4. Потенциальные риски	

Дата выдачи задания для раздела по линейному графику	01.03.2020

Задание выдал консультант:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОСГН ШБИП	Меньшикова Екатерина Валентиновна	к.ф.н		02.06.2020

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
8ПМ8И	Журбич Никита Игоревич		02.06.2020

ЗАДАНИЕ ДЛЯ РАЗДЕЛА
«СОЦИАЛЬНАЯ ОТВЕТСТВЕННОСТЬ»

Студенту:

Группа	ФИО
8ПМ8И	Журбич Никита Игоревич

Школа	ИШИТР	Отделение (НОЦ)	ОИТ
Уровень образования	Магистратура	Направление/специальность	09.04.04 Программная инженерия

Тема ВКР:

Методология подготовки исходных данных для модели машинного обучения в нефтегазовой области	
Исходные данные к разделу «Социальная ответственность»:	
1. Характеристика объекта исследования (вещество, материал, прибор, алгоритм, методика, рабочая зона) и области его применения	**Объект исследования** – методология подготовки исходных данных с нефтегазового месторождения для дальнейшего построения модели машинного обучения. **Рабочая зона** - аудитория, оборудованная системой отопления, кондиционирования воздуха, с естественным и искусственным освещением. Рабочее место – стационарное, оборудованное персональным компьютером и оргтехникой.
Перечень вопросов, подлежащих исследованию, проектированию и разработке:	
1. Правовые и организационные вопросы обеспечения безопасности. 1.1.Специальные (характерные при эксплуатации объекта исследования, проектируемой рабочей зоны) правовые нормы трудового законодательства. 1.2.Организационные мероприятия при компоновке рабочей зоны.	• Трудовой кодекс Российской Федерации от 30.12.2001 N 197-ФЗ • Федеральный закон от 27.07.2006 N 152-ФЗ (ред. От 25.07.2011) «О персональных данных» • ГОСТ 12.2.032-78 ССБТ
2. Производственная безопасность. 2.1. Анализ вредных и опасных факторов, которые может создать объект исследования. 2.2. Анализ вредных и опасных факторов, которые могут возникнуть на рабочем месте. 2.3. Обоснование мероприятий по защите исследователя от действия опасных и вредных факторов.	Микроклимат: • СанПин 2.2.4.548-96 • СанПин 2.2.2/2.4.1340-03 Шум: • ГОСТ 12.1.003-2014 ССБТ • ГОСТ 12.1.029-80 ССБТ Освещённость, естественный и искусственный свет: • СанПиН 2.2.1/2.1.1.1278-03 Поражение электрическим током: • ГОСТ 12.1.019-2017

	Умственное перенапряжение: • ТОИ Р-45-084-01
3. Экологическая безопасность. 3.1. Анализ влияния объекта исследования на окружающую среду. 3.2. Анализ влияния процесса исследования на окружающую среду 3.3. Обоснование мероприятий по защите окружающей среды	Несущими угрозу окружающей среде являются следующие объекты: • Люминесцентные лампы • Компьютерное оборудование Утилизацию следует проводить согласно: • ГОСТ Р 56397-205 • ГОСТ 12.3.031-83
4. Безопасность в чрезвычайных ситуациях. 4.1. Анализ вероятных ЧС, которые может инициировать объект исследований. 4.2. Анализ вероятных ЧС, которые могут возникнуть на рабочем месте при проведении исследований. 4.3. Обоснование мероприятий по предотвращению ЧС и разработка порядка действия в случае возникновения ЧС.	Наиболее вероятной ЧС является пожар в здании, в т.ч. из-за короткого замыкания электропроводки. Необходимо принять ряд предупреждающих мер. В случае возникновения ЧС вызвать противопожарную службу, эвакуировать людей и, согласно плану эвакуации, покинуть помещению. Для тушения локальных очагов возгорания возможно использование огнетушителей типа ОУ-5.

Дата выдачи задания для раздела по линейному графику	**01.03.2020**

Задание выдал консультант:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ООД ШБИП	Горбенко Михаил Владимирович	к.т.н		10.06.2020

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
8ПМ8И	Журбич Никита Игоревич		10.06.2020

Планируемые результаты обучения

Код	Результат обучения
Общие по направлению 09.04.04 «Программная инженерия»	
Р1	Проводить научные исследования, связанные с объектами профессиональной деятельности
Р2	Разрабатывать новые и улучшать существующие методы и алгоритмы обработки данных в информационно-вычислительных системах
Р3	Составлять отчеты о проведенной научно-исследовательской работе и публиковать научные результаты
Р4	Проектировать системы с параллельной обработкой данных и высокопроизводительные системы
Р5	Осуществлять программную реализацию информационно-вычислительных систем, в том числе распределенных
Р6	Осуществлять программную реализацию систем с параллельной обработкой данных и высокопроизводительных систем
Р7	Организовывать промышленное тестирование создаваемого программного обеспечения
Профиль «Технологии больших данных»/ «Big data solutions»	
Р8	Исследовать и анализировать большие данные, создавать их модели и интерпретировать структуры данных в таких моделях
Р9	Понимать принципы создания, хранения, управления, передачи и анализа больших данных с использованием новейших технологий, инструментов и систем обработки данных в высокопроизводительных сетях
Р10	Применять теорию распределенной системы управления базами данных к традиционным распределенным системам реляционных баз данных, облачным базам данных, крупномасштабным системам машинного обучения и хранилищам данных

Реферат

Выпускная квалификационная работа 107 с., 21 рис., 23 табл., 26 источников, 4 прил.

Ключевые слова: методология обработки данных, разведочный анализ, регрессионный анализ, набор данных, очистка данных.

Объектом исследования является процесс разработки методологии обработки исходных данных с нефтегазового месторождения для построения прогнозной модели.

Цель работы – формализация и разработка методологии обработки данных для построения модели машинного обучения.

В процессе исследования был проведен анализ основных проблем в рассматриваемой области, поставлены цели для их непосредственного выполнения. Также был проведен разведочный анализ данных для построения модели машинного обучения. После этого был проведен регрессионный анализ и построение модели.

В результате исследования были формализованы требования к разработке методологии обработки данных. Разработана методология обработки данных и модель машинного обучения для прогнозирования определённых параметров из набора данных.

Степень внедрения: на текущем этапе внедрение не планируется.

Область применения: исследовательские и производственные проекты в нефтегазовой промышленности

Экономическая эффективность/значимость работы данная разработка облегчает задачу по обработке данных из нефтегазового сектора, сокращая количество атрибутов. Это приводит к экономии ресурсов памяти, что в реалиях обработки больших данных имеет большое значение.

В будущем планируется разработка других типов машинного обучения (например модель кластеризации).

Оглавление

Перечень условных обозначений и терминов

- Big Data («большие данные») - обозначение структурированных и неструктурированных данных огромных объёмов и значительного многообразия, эффективно обрабатываемых горизонтально масштабируемыми программными инструментами.
- Автомобильная заправочная станция – АЗС
- Бассейновое моделирование - это динамический анализ, в основе которого лежит численное моделирование геологических процессов, протекающих в осадочных бассейнах.
- Фонд скважин — число и классификация по состоянию и назначению всех пробуренных скважин (на месторождении, газовом промысле или подземном хранилище газа). В этот фонд входят все разведочные, эксплуатационные, наблюдательные и специальные скважины.
- Геолого-технические мероприятия – ГТМ
- Пласт - слой почвы, осадочной или магматической породы, который сформировался на поверхности Земли и имеет внутреннюю структуру, которая отличается от других слоёв, лежащих непосредственно над ними и лежащих под ними по цвету, текстуре, материалу.
- Электроцентробежный насос – ЭЦН
- Выброс (англ. outlier) - в статистике результат измерения, выделяющийся из общей выборки.
- Обводнённость скважины - содержание воды в продукции скважины, определяемое как отношение дебита воды к сумме дебитов нефти и воды.
- Дебит - объём жидкости (воды, нефти) или газа, стабильно поступающий из некоторого естественного или искусственного источника в единицу времени.
- Диаметр насосно-компрессорных труб – диаметр НКТ
- CSV (от англ. Comma-Separated Values - значения, разделённые запятыми) - текстовый формат, предназначенный для представления

табличных данных. Строка таблицы соответствует строке текста, которая содержит одно или несколько полей, разделенных запятыми.

- Попутный газ - газ, растворенный в нефти.
- Газовая шапка - скопление свободного газа в наиболее приподнятой части нефтяного пласта, над нефтяной залежью.
- Фонтанный способ добычи нефти – ФОН
- KNN (k-nearest algorithm imputation) – алгоритм поиска ближайшего соседа для восстановления пропущенных значений.
- Training samples – тренировочная выборка
- Testing samples – тестовая выборка
- Нормальное распределение, также называемое распределением Гаусса или Гаусса - Лапласа - распределение вероятностей, которое в одномерном случае задаётся функцией плотности вероятности, совпадающей с функцией Гаусса.

Введение

Нефтегазовые компании в процессе своей деятельности получают петабайты данных каждый день, использование больших данных открывает возможности анализа и предсказания развития трендов в области геологии, инженерии, производства и наилучшего способа использования оборудования для достижения наиболее оптимальных результатов работы на всех стадиях своей деятельности.

Нефтяные и газовые компании не смогут воспользоваться конкурентным преимуществом технологий Big Data, если не начнут более эффективно управлять своими данными. К такому выводу в своем новом докладе пришла нефтегазовая консалтинговая компания Molten. По мнению ее экспертов, многие нефтегазовые предприятия «безответственно» распоряжаются своими данными, несмотря на то, что тратят миллиарды долларов в год на их сбор. По подсчетам Molten, крупные нефтегазовые компании тратят от $1 до $3 млрд в год на сбор данных, однако расходы на поддержание и обработку накопленной информации зачастую составляют менее 1% от этой суммы. В то же время от компаний требуется принимать оперативные решения и поддерживать высокий уровень производительности. Как следствие, руководство должно полагаться на большие объемы данных, чтобы принимать критические решения. Сфера применения технологии Big Data в нефтегазовой отрасли очень обширна, и включает весь спектр, от геологоразведки и разработки до переработки углеводородного сырья [1].

Целью данной магистерской диссертации является разработка методологии обработки исходных данных с нефтяного месторождения для построения прогнозной модели.

1. Обзор предметной области

1.1. Области применения технологий Big Data в нефтяном инжиниринге

В настоящее время технологии Big Data являются одним из ключевых драйверов развития информационных технологий. Однако успешных кейсов в мировом нефтяном инжиниринге немного. Это связано с характерной для многих фундаментальных отраслей исторически-наследованной инерционностью. На протяжении длившегося около 10 лет периода высоких цен на нефть информационные технологии не рассматривались как значимый драйвер роста в нефтяной промышленности. Современные реалии диктуют необходимость оптимизации процессов и повышения операционной эффективности в нефтегазовом секторе.

Большое число успешных проектов реализовано в области автоматизации обработки данных, например, в проектах по созданию цифровых месторождений, и в предиктивной аналитике для оценки надежности и прогнозирования осложнений при эксплуатации оборудования в различных технологических процессах, преимущественно в бурении. По оценочным расчетам, внедрение систем предиктивной аналитики на основе анализа больших данных в бурении позволяет сократить сроки строительства скважин на 30 %, а общую стоимость скважины, включая освоение, на 15 %. Также посредством инструментов Big Data успешно решается широкий круг задач в логистике: от оптимизация транспортных маршрутов и схем поставок оборудования до повышения эффективности работы АЗС [2].

На рисунке 1 были выделены приоритеты внедрения технологий Big Data в нефтегазовом секторе.

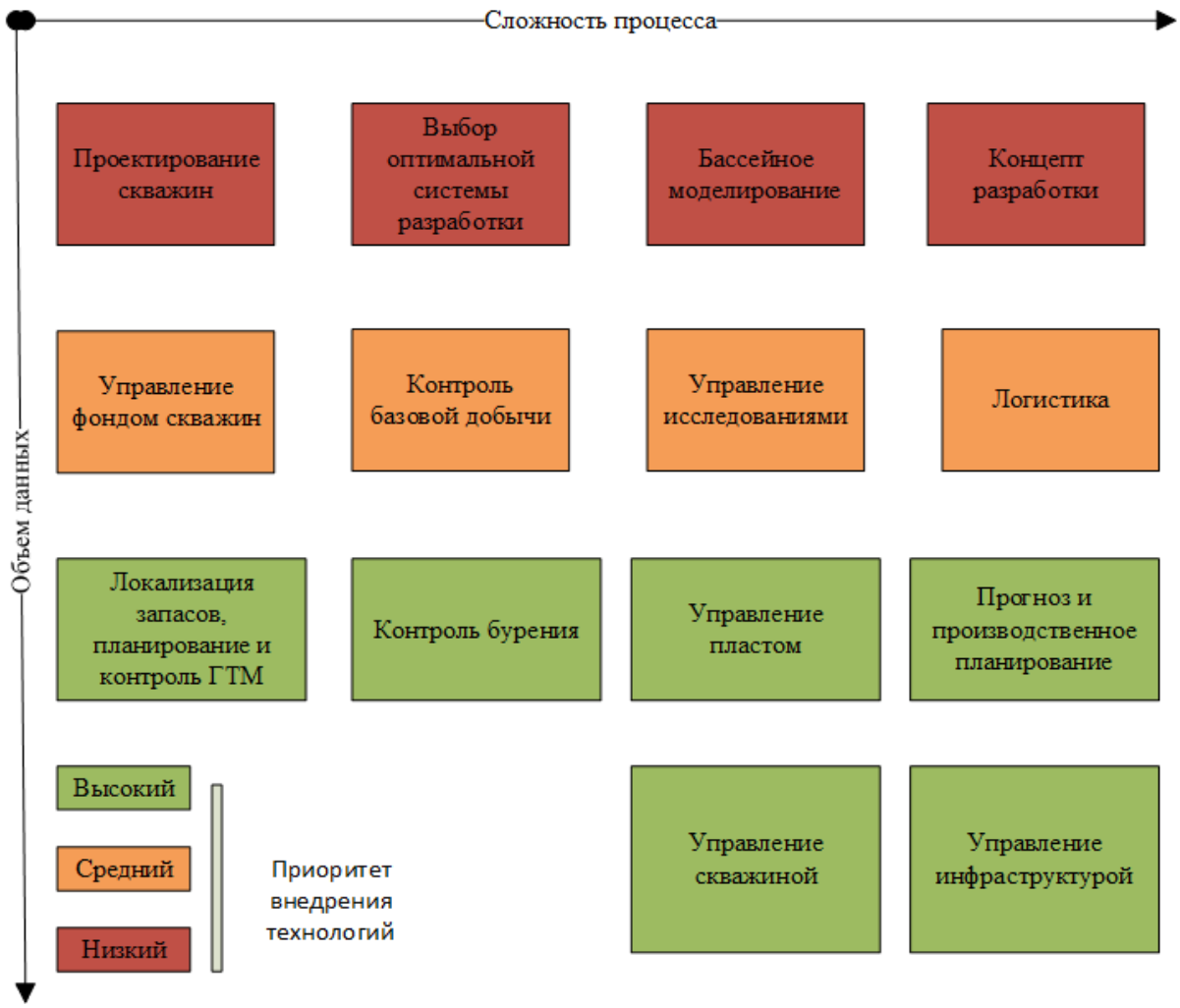


Рисунок 1. Приоритеты внедрения технологий Big Data в нефтегазовом секторе

Исходя из представленных приоритетов внедрения технологий Big Data в нефтегазовом секторе можно отметить самые приоритетные задачи: управление скважиной, управление инфраструктурой, прогноз и производственное планирование, управление пластом. В данной работе одной из приоритетных задач будет создание модели машинного обучения для прогнозирования определённых параметров.

1.2. Проблема, цель

Для изучения и поиска причин рассматриваемой проблемы используется диаграмма «Рыбий скелет». Данная диаграмма позволяет в простой и доступной форме определить все потенциальные причины рассматриваемой проблемы.

В качестве основной проблемы выбрана следующая проблема: «Отсутствие аналогов подобной системы на рынке». На рисунке 2 продемонстрирована диаграмма «Рыбий скелет» для анализа предметной области и поиска причины.

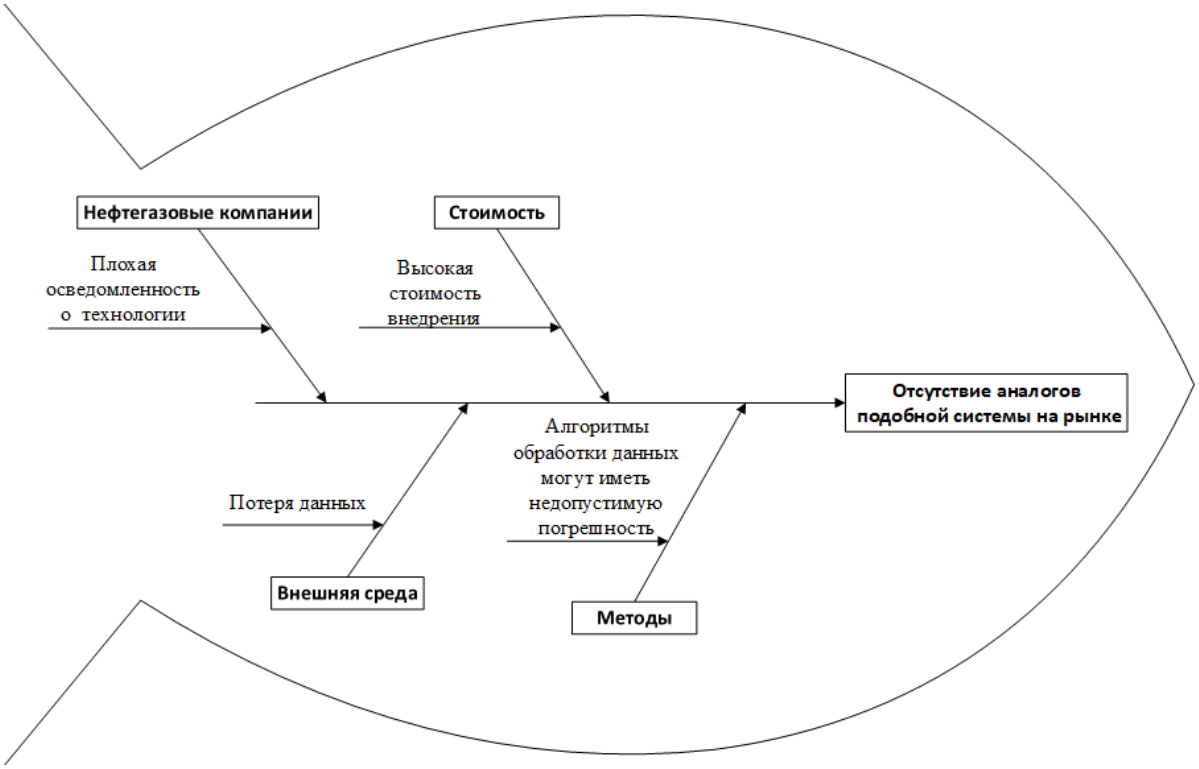


Рисунок 2. Диаграмма «Рыбий скелет» для рассматриваемой предметной области

Проблема - отсутствие аналогов рассматриваемой методологии на рынке. Большинство нефтегазовых компаний только сейчас приступают к использованию инструментов анализа больших данных, не смотря на успех этих инструментов в других областях производства.

Цель - разработка методологии, которая позволит проводить корректный разведочный анализ данных.

1.3. Опыт по использованию технологий Big Data в нефтегазовой отрасли

Зарубежный опыт по использованию технологий Big Data в нефтегазовой отрасли:

1. *Сокращение сроков строительства скважин на 30 %, снижение общей стоимости скважины — на 15%.*

Британской компании British Petroleum удалось снизить затраты по эксплуатации оборудования более чем на 2 миллиона долларов благодаря системам прогнозирования скважин электроцентробежных насосов на морских платформах. Это способствовало повышению периода работы скважин и уменьшению времени простоя в ожидании ремонта оборудования.

2. *Оптимизация расписания диагностических проверок нефтегазового оборудования.*

Американская компания GE, General Electric Oil & Gas (специализируется на производстве оборудования для нефтегазовой области) произвела внедрение алгоритмов машинного обучения в одно из своих подразделений. Автоматизированный анализ собранных данных позволял инженерам компании оптимизировать расписание диагностических проверок, улучшить эффективность использования оборудования и снизить время «простоя» за счет превентивного выявления возможных неисправностей.

3. *Снижение себестоимости добычи нефтепродуктов за счет концепции «цифрового месторождения».*

Цифровым месторождением называют активы, которые оснащены набором систем мониторинга и удаленного контроля, а также специализированным программным обеспечением для производственных процессов. Подход нацелен на прирост добычи нефти и газа, а также сокращение простоев и трудозатрат за счет оптимизации работ и снижения недоборов. Цифровые месторождения обеспечивают оптимальный технологический режим добычи нефти, что позволяет снизить себестоимость

добычи на 7-10%, а себестоимость эксплуатации промыслового объекта – на 20%.

Отечественный опыт использования технологий Big Data в нефтегазовой отрасли:

1. Выявление причин сбоев автоматического перезапуска насосов после аварийного отключения электропитания.

В ПАО «Газпромнефть» были проанализированы более 200 миллионов различных записей, полученных за год с контроллеров систем управления на 1649 скважинах, записи рестартов напряжения из аварийных журналов и факторы зависимости работы насосов от скважинных условий, особенностей эксплуатации, схемы электроснабжения и др.

Аналитические инструменты позволили сформировать и проверить набор гипотез о причинах сбоев и получить информацию о ранее неизвестных взаимосвязях в работе насосного оборудования, например, о появлении эффекта турбинного вращения, приводящего к обратному сливу нефти при отключении электропитания насоса.

2. Выявление воровства газа.

В 2018 г. «Газпром» разработал новый аналитический алгоритм для установки во всех своих дочерних газоснабжающих и газораспределительных компаниях. Например, сейчас наиболее остро проблема воровства стоит в Северо-Кавказском федеральном округе, где потери газа достигают 3,5 миллиарда кубометров в год, т.е. 16 миллиардов рублей. Планируется сократить дисбаланс учтенного и неучтенного топлива за счет непрерывного сбора данных о потреблении, их систематизации и мониторинга потребления по балансовым зонам. Прогнозируемая отслеживаемость – не менее 90% от всего объема потребления [3].

1.4. Преимущества и недостатки разработки методологии

Обработка больших данных меняет практически все сферы нашей жизни, и бизнес в первую очередь. Но пока еще почему-то не все предприятия, даже крупные, перешли на использование этой технологии. А если и используют ее, то не в полной мере. Соответственно, у данной технологии есть свои недостатки.

В таблице 1 приведены основные преимущества и недостатки разработки методологии для обработки данных, полученных с нефтегазового месторождения.

Таблица 1. Основные преимущества и недостатки методологии обработки данных

Преимущества	Недостатки
Минимизации влияния «человеческого фактора» при интерпретации исследований	Повышенные требования к количеству и качеству входных данных для сложных моделей.
Повышение качества и своевременности принятия производственных решений	Большие трудозатраты на создание модели и обработку результатов
Организация неструктурированной информации, состоящей из текстов, изображений, видео и других типов данных	Проблема выбора обрабатываемых данных: то есть определение того, какие данные необходимо извлекать, хранить и анализировать

Несмотря на повышенные требования к программному и аппаратному обеспечению и большим затратам на хранение и обработку данных, самым большим недостатком является проблема выбора обрабатываемых данных: необходимо определить, какие данные необходимо извлекать, хранить и анализировать, а какие данные нужно удалить.

1.5. Вывод по разделу

Проблема обработки и хранения больших данных в нефтегазовых компаниях остается актуальной и по сей день в силу ряда причин. В первую очередь, это невозможность использования традиционных подходов. Несмотря на разнообразие технологий и способов решения данной задачи, нет универсального способа обработки большого количества данных, полученных с нефтегазовых месторождений. Данный вывод подчеркивает необходимость создания методологии обработки этих данных для дальнейших исследований и проведения экспериментов в нефтегазовом секторе.

2. Разведочный анализ данных

2.1. Выбор инструментов разработки

Подготовка и чистка исследуемого датасета для дальнейшей визуализации производилась на языках программирования Python и R.

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. Предоставляет специальные структуры данных и операции для манипулирования числовыми таблицами и временными рядами.
- Seaborn - это библиотека визуализации данных Python, основанная на matplotlib. Она предоставляет высокоуровневый интерфейс для рисования привлекательной и информативной статистической графики.
- R Studio — свободная среда разработки программного обеспечения с открытым исходным кодом для языка программирования R, который предназначен для статистической обработки данных и работы с графической репрезентацией.

2.2. Очистка исходных данных

Исследуемый файл содержит информацию о нефтегазовом месторождении, основных параметрах и показателях рассматриваемого месторождения. Файл содержит 3368 записей и более 120 атрибутов.

Для начала необходимо посмотреть на структуру набора данных. Структура файла представлена на рисунке 3.

	Скважина	ГТМ	Метод	Состояние	Время работы, ч	Простой, ч	Причина простоя	Обводненность (вес), %	Нефть, м3
0	53514b4c4150ad897d82dd7d42cfc1a5	0.0	ЭЦН/ ФОН	РАБ.	301.0	0.0	NaN	69.7	320.17
1	53514b4c4150ad897d82dd7d42cfc1a5	0.0	ЭЦН	РАБ.	720.0	0.0	NaN	38.2	1235.33
2	53514b4c4150ad897d82dd7d42cfc1a5	0.0	ЭЦН	РАБ.	744.0	0.0	NaN	35.1	1203.22
3	53514b4c4150ad897d82dd7d42cfc1a5	0.0	ЭЦН	РАБ.	744.0	0.0	NaN	32.1	1035.57
4	53514b4c4150ad897d82dd7d42cfc1a5	0.0	ЭЦН	РАБ.	720.0	0.0	NaN	34.2	875.54

Рисунок 3. Структура рассматриваемого набора данных

Прежде всего, нужно подготовить данные для анализа. Для этого нужно понять, есть ли в наборе данных отсутствующие или нулевые значения. Наша цель - изменить значения в наборе данных, если это возможно. Это очень важный этап анализа данных, потому что если в наборе данных содержатся некорректные данные на входе, будут плохие результаты в конце.

На данном этапе работы было предложено решить следующие задачи:

- Устранение ошибок (errors) и восстановление пропущенных значений (missing values) в наборе данных
- Определение выбросов (outliers) с помощью графических представлений нескольких атрибутов (признаков)
- Проверка атрибутов на мультиколлинеарность

В таблице 2 представлены основные этапы чистки данных [4].

Таблица 2. Этапы чистки данных

Исходные («грязные») данные	Формат переменных	Предполагаемые действия /корректировки/
1.*Missing data*/ Отсутствующие данные	**Numeric**/числовой, **Char**/текст	**1. Add in** (average, median, frequency...) /1. Заменить (средним, медианой, частотой...)/ **2.Delete this cases** (rows) /2.Удалить эти записи/
2.*Mistakes of data*/ Ошибки в данных	**Numeric**/числовой, **Char**/текст	**1. Add in** (average, median, frequency...) /1. Заменить (средним, медианой, частотой...)/ **2.Delete this cases** (rows) /2.Удалить эти записи/
3.*Outliers of data* / Выбросы данных	**Numeric**/числовой	**Delete this cases** (rows) /Удалить эти записи/
4.*Duplicate cases*(rows) /Дублирующие наблюдения(строки)	**Duplicate ID** (observations)	**Remove one of the duplicate** /Убрать одну из дублирующих записей/
5.*Multicollinearity in the original data*/ Мультиколлинеарность	**Linear combination of variables** (attributes)	**Remove one of the attributes/** /Убрать один из атрибутов/

На рисунке 4 проиллюстрированы общие статистические данные для рассматриваемого набора данных.

	ГТМ	Время работы, ч	Закачка, м3	Газ из газовой шапки, м3	Дебит конденсата	Диаметр экспл.колонны	Диаметр НКТ	Диаметр штуцера	Глубина верхних дыр перфорации
count	2906.000000	2906.000000	2906.000000	2906.0	2896.0	1239.000000	1239.000000	1229.000000	1239.000000
mean	144.999656	571.711287	198.424295	0.0	0.0	143.364052	60.510412	10.317331	3099.808055
std	2465.153189	274.318835	2514.230449	0.0	0.0	35.094014	24.674615	14.384388	583.190372
min	0.000000	0.000000	0.000000	0.0	0.0	0.000000	0.000000	0.000000	0.000000
25%	0.000000	505.750000	0.000000	0.0	0.0	146.800000	67.300000	0.000000	2902.000000
50%	0.000000	720.000000	0.000000	0.0	0.0	146.800000	67.900000	0.000000	3230.010000
75%	0.000000	744.000000	0.000000	0.0	0.0	159.600000	73.000000	32.000000	3445.000000
max	42438.000000	744.000000	42436.000000	0.0	0.0	159.600000	89.000000	32.000000	4141.000000

Рисунок 4. Общие статистические данные

Затем нужно проанализировать, какие типы данных встречаются в исследуемом наборе данных. На рисунке 5, после чтения CSV-файла, продемонстрирована проверка имен столбцов и атрибутов с помощью метода библиотеки pandas.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3367 entries, 0 to 3366
Columns: 123 entries, Скважина to ГП - Общий прирост Qн
dtypes: float64(66), object(57)
memory usage: 2.4+ MB
```

Рисунок 5. Типы встречающихся данных

На рисунке 6 можно увидеть, какие столбцы имеют пропущенные значения или нулевые значения.

```
Скважина                  False
Дата                      False
ГТМ                        True
Метод                      True
Характер работы            True
Состояние                  True
Время работы, ч            True
Время накопления           True
Нефть, т                   True
Попутный газ, м3           True
Закачка, м3                True
Природный газ, м3          True
```

Рисунок 6. Проверка на наличие пропущенных значений

Для восстановления пропущенных значений в данном случае будет использоваться пакет R Studio с языком программирования R.

Прежде всего необходимо считать файл и заменить пустые или пропущенные значения на пустые (NA). На рисунке 7 продемонстрирован данный процесс.

	Скважина	Дата	ГТМ	Метод	Характер.работы	Состояние	Время.работы..ч
1	002ff5b8a6dc271f58581e1b4fa2c5fc	01.12.2016	1	ФОН	НЕФ	ОСВ ТГ	0
2	008d0347e572a5d938a9c40c29e539fc	01.10.2013	NA	<NA>	<NA>	<NA>	NA
3	00b40cb7bb8c9fd1ac26b4cc86f2b291	01.02.2018	NA	<NA>	<NA>	<NA>	NA
4	01ba18d8b6d29875a18d4bca4eb201d7	01.05.2014	0	ЭЦН/ФОН	НЕФ	РАБ.	120
5	024ec6f6e3f9c5150ecf525bf8b7a6a3	01.06.2017	1	ФОН	НЕФ	ОСВ ТГ	0
6	0254a227c6c2c31a419126700cfcddc2	01.05.2017	1	ЭЦН/ФОН	НЕФ	ОСТ.	193

Рисунок 7. Чтение файла в среде разработки R Studio

После этого появляется возможность просмотреть количество пустых значений (NA). На рисунках 8 и 9 приведено подробное описание определённых атрибутов.

Variables sorted by number of missings:

Variable	Count
Причина.простоя	215
ГТМ	45
Метод	45
Характер.работы	45
Состояние	45
Время.работы..ч	45
Попутный.газ..м3	45
Простой..ч	45
Обводненность..вес....	45
Добыча.растворенного.газа..м3	45
Дебит.попутного.газа..м3.сут	45
Скважина	0
Дата	0

Рисунок 8. Количество пропущенных значений в каждом атрибуте

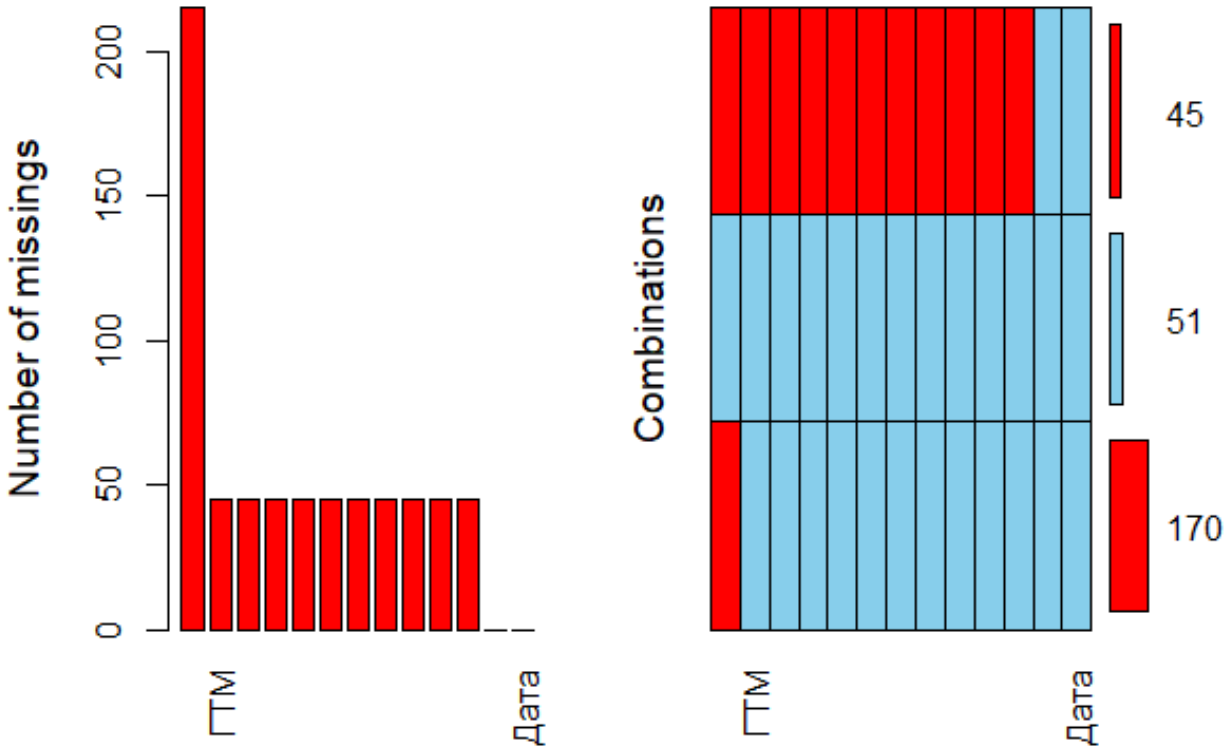


Рисунок 9. Количество пропущенных значений в каждом атрибуте

Исходя из приведённых графических представлений можно сделать вывод о том, что атрибут «Причина простоя» имеет наибольшее количество пропущенных значений. Также атрибуты «Скважина» и «Дата» не имеют пропущенных значений совсем.

2.3. Определение выбросов и ошибок данных

После подготовки данных исследуемого датасета можно приступить к визуализации данных. Например, на рисунке 10 можно увидеть самые распространенные методы, используемые на буровых скважинах. Первое место – это ЭЦН (электроприводной центробежный насос) - наиболее широко распространённый в России аппарат механизированной добычи нефти [5]. Второе место – комбинация ЭЦН и ФОН (фонтанный способ добычи нефти). Третье место - ФОН (фонтанный способ добычи нефти) [6].

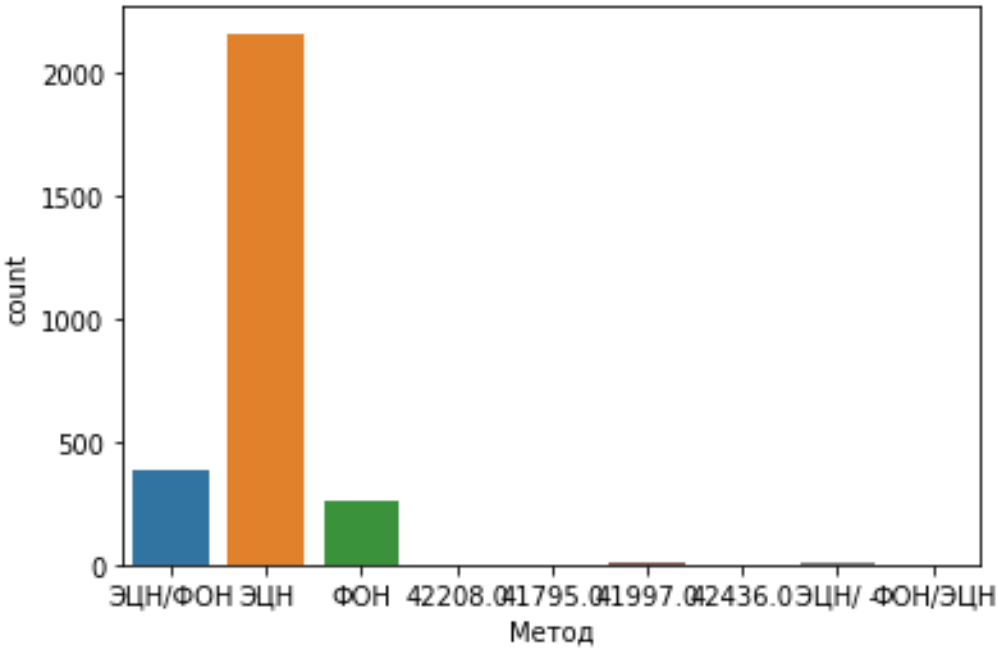


Рисунок 10. Методы, используемые на буровых скважинах

На рисунке 11 показано наличие проведения ГТМ (Геолого-технические мероприятия – это работы, проводимые на скважинах с целью регулирования разработки месторождений и поддержания целевых уровней добычи нефти) [7].

В данном случае используется словарь данных:

- Значение «1» - геолого-технические мероприятия проводились на данной скважине;
- Значение «0» - геолого-технические мероприятия не проводились на данной скважине.

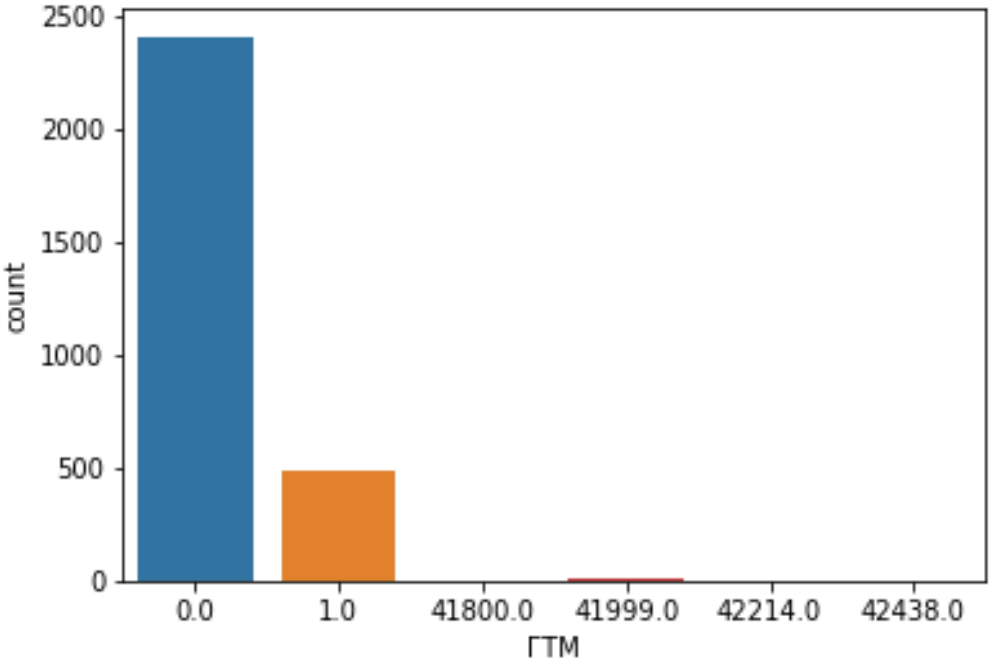


Рисунок 11. Наличие проведения ГТМ

Одно из главных преимуществ визуализации перед другими методами является возможность графически представить большое количество информации и понять, какие ошибки могли появиться после этапа подготовки данных для дальнейшего анализа. В данном случае можно заметить, что в данном наборе данных присутствуют значения, которые выбиваются из диапазона данных (выбросы). Также при графическом представлении можно легко заметить ошибочные данные (например формат дат при искомом числовом варианте). Следовательно, рекомендуется удалить такие данные, так как в будущем они могут существенно повлиять на проведение регрессионного анализа.

2.4. Кодирование категориальных признаков

Подавляющее большинство методов классификации и регрессии сформулированы в терминах евклидовых или метрических пространств, то есть подразумевают представление данных в виде вещественных векторов одинаковой размерности. В реальных данных, однако, не так редки категориальные признаки, принимающие дискретные значения. Определим то, как работать с такими данными, в частности с помощью линейных моделей, и что делать, если категориальных признаков много, да еще и у каждого большое количество уникальных значений.

С помощью библиотеки Sklearn использован класс LabelEncoder для кодирования категориальных признаков:

```
from sklearn.preprocessing import LabelEncoder
le_con = LabelEncoder()
le_con.fit(df_filt['Метод'])
df_filt['Метод'] = le_con.fit_transform(df_filt['Метод'].values)
```

Рисунок 12. Кодирование категориального признака «Метод»

Метод fit этого класса находит все уникальные значения и строит таблицу для соответствия каждой категории некоторому числу, а метод transform непосредственно преобразует значения в числа. После fit у label_encoder будет доступно поле classes_, содержащее все уникальные значения. Можно их пронумеровать и убедиться, что преобразование выполнено верно.

2.5. Восстановление пропущенных значений

Первым шагом при восстановлении пропущенных значений является проверка атрибутов на нормальность распределения. Этот шаг изображен на рисунке 13. В данном случае проверку проходит атрибуты «Нефть» и «Дебит жидкости».

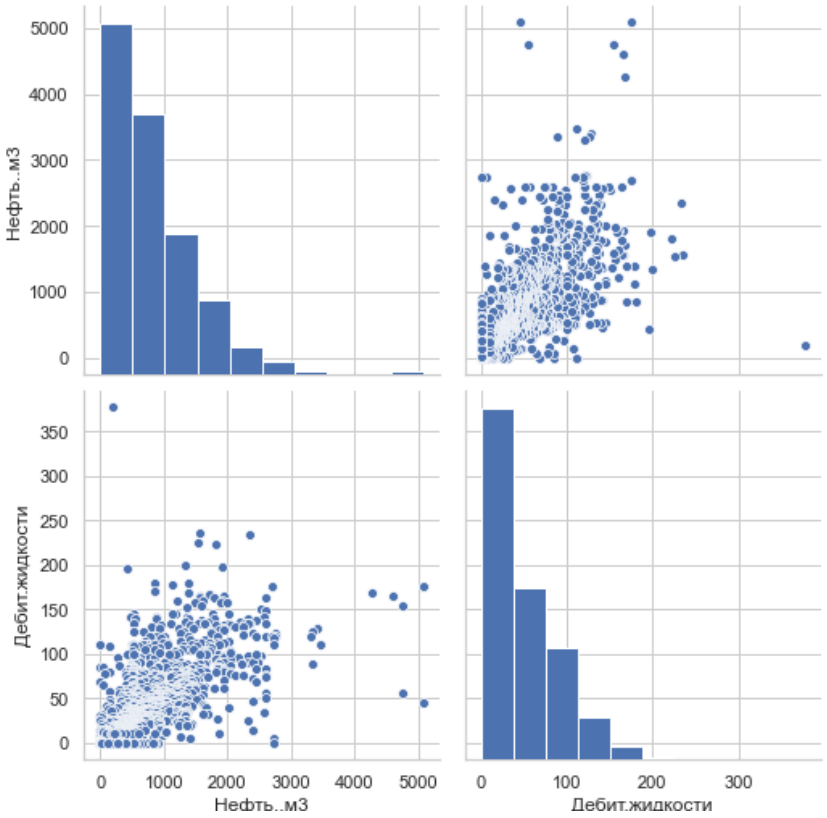


Рисунок 13. Проверка атрибутов на нормальность распределения

Исходя из результатов проверки можно сделать вывод о том, что атрибуты «Нефть» и «Дебит жидкости» успешно прошли проверку на нормальность распределения.

Следующим шагом является непосредственно восстановление пропущенных значений. Для восстановление пропущенных значений в данном наборе данных был выбран алгоритм поиска ближайшего соседа (k-nearest algorithm imputation).

KNN - это алгоритм, который полезен для сопоставления точки с ее ближайшими k соседями в многомерном пространстве. Он может использоваться для данных, которые являются непрерывными, дискретными, порядковыми и категориальными, что делает его особенно полезным для работы со всеми видами недостающих данных.

Предположение за использование KNN для пропущенных значений состоит в том, что значение точки может быть аппроксимировано значениями ближайших к нему точек на основе других переменных [8].

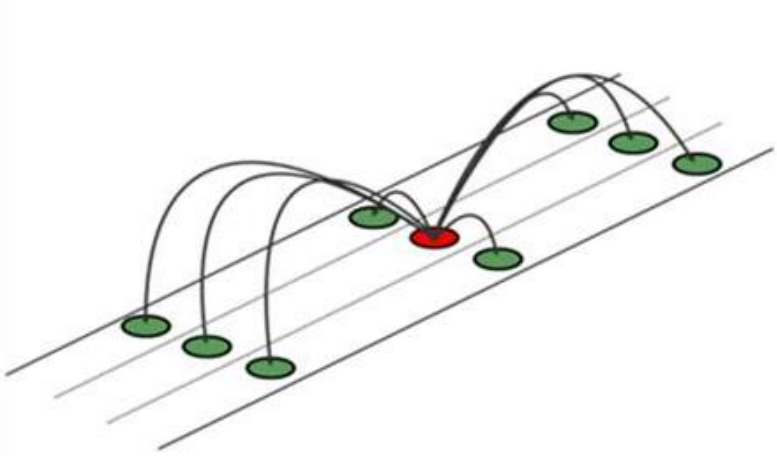


Рисунок 14. Принцип работы алгоритма поиска ближайшего соседа

После применения данного алгоритма вставки пропущенных значений, можно проверить переменные. Проверка корректности работы алгоритма поиска ближайшего соседа (k-nearest algorithm imputation) представлена на рисунках 15 и 16.

Variables sorted by number of missings:

Variable	Count
Скважина	0
Дата	0
ГТМ	0
Метод	0
Характер.работы	0
Состояние	0
Время.работы..ч	0
Попутный.газ..м3	0
Простой..ч	0
Причина.простоя	0
Обводненность..вес....	0
Добыча.растворенного.газа..м3	0
Дебит.попутного.газа..м3.сут	0

Рисунок 15. Результат работы алгоритма поиска ближайшего соседа

	Скважина	Метод	Состояние	Попутный.газ..м3	Нефть..м3	Жидкость..м3	Добыча.растворенного.газа..м3
0	300d3d8ef824d7963f0eb362908ff183	ЭЦН	РАБ.	136280	1598.44	2924.91	136279.53
1	64fcc93f7557e22482c171cac2e07d82	ЭЦН	РАБ.	52847	1845.78	3238.08	52847.27
2	d3b63715faebec4d81754c3e7a3ad836	ЭЦН	РАБ.	77725	2539.15	2937.34	56808.00
3	5d995ad52dbe6eff8f1b6681d6b21bec	ЭЦН	РАБ.	58760	758.38	1422.56	58760.12
4	ef3f72b14c3febb4d177c5e92ae4b561	ЭЦН	РАБ.	122280	2332.96	2271.86	121011.44
5	429ecc93c8d72b9645bfd48c0f826dc7	ЭЦН	РАБ.	139720	1581.82	2548.04	139720.36
6	5756c913481872253c810e989328d7a5	ЭЦН	РАБ.	132117	1944.00	2816.00	132117.00
7	475a517d2c9ac61bd91a386b24824610	ЭЦН	РАБ.	148053	1706.47	2837.51	148053.11

Рисунок 16. Результат работы алгоритма поиска ближайшего соседа

Исходя из результатов проверки можно сделать вывод о том, что в ходе работы данного алгоритма все значения были успешно восстановлены. Также после восстановления пропущенных значений необходимо убедиться в том, что все атрибуты подчиняются нормальному закону распределения. На рисунке 17 представлен пример успешной проверки нескольких атрибутов на нормальность распределения.

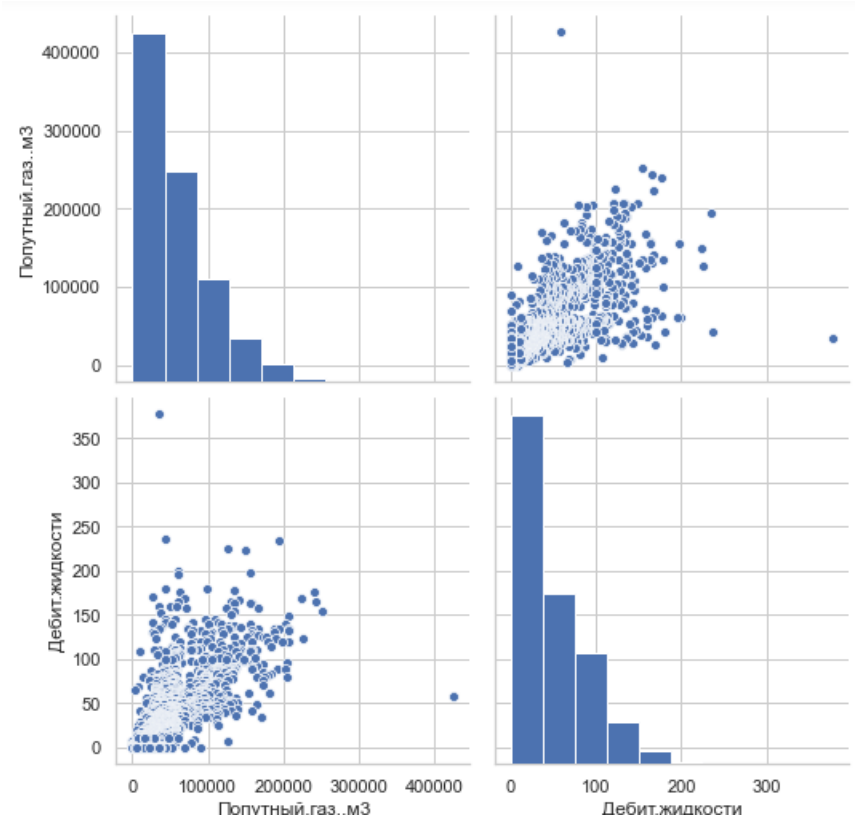


Рисунок 17. Проверка атрибутов на нормальность распределения после восстановления значений

2.6. Проверка на мультиколлинеарность

Мультиколлинеарность (multicollinearity) - наличие линейной зависимости между объясняющими переменными (факторами) регрессионной модели. При этом различают полную коллинеарность, которая означает наличие функциональной (тождественной) линейной зависимости и частичную или просто мультиколлинеарность — наличие сильной корреляции между факторами.

На следующем рисунке представлен небольшой отрезок матрицы факторов для проверки мультиколлинеарности.

Рисунок 18. Матрица факторов

Индикатор мультиколлинеарности: в корреляционной матрице встречаются элементы, по модулю близкие к 1. Таких элементов в данном наборе данных было несколько, например атрибуты «Нефть, м3» и «Нефть, т» линейно зависят друг от друга, поэтому один из данных атрибутов было решено удалить из рассматриваемой выборки.

Благодаря построению данной матрицы удалось исключить порядка 20 атрибутов, которые линейно зависимы от целевой функции.

2.7. Вывод по разделу

В результате проведения разведочного анализа данных были решены следующие задачи:

- Определены и устранены выбросы в данных;
- Отсутствующие данные успешно восстановлены;
- Удалены дубликаты и пустые строки;
- Удалены ошибки в данных;
- Категориальные данные закодированы;
- Все данные проверены на мультиколлинеарность

Благодаря проведению разведочного анализа удалось исключить порядка 50 атрибутов и 500 строк, которые линейно зависимы от целевой функции. Также после корректного проведения разведочного анализа возможно проведение регрессионного анализа.

3. Регрессионный анализ

3.1. Выбор целевой функции

Целевая функция есть математическое выражение некоторого критерия качества одного объекта (решения, модели, процесса и т.д.) в сравнении с другим.

Целевая функция может быть записана так:

$$q(x_{0оц}, x_{1оц}, \dots, x_{Nоц}) = \sum_{i=1}^{N}(xi - xi\ оц)\sum_{i=1}^{N}(xi - xiоц)^2, \quad (1)$$

Цель – найти такие оценки Xiоц, при которых целевая функция достигает минимума.

В данной работе целевой функцией будет выбран параметр «Нефть, м3».

	Скважина	Метод	Состояние	Попутный.газ..м3	Нефть..м3	Жидкость..м3
2	300d3d8ef824d7963f0eb362908ff183	ЭЦН	РАБ.	136280	1598.44	2924.91
3	64fcc93f7557e22482c171cac2e07d82	ЭЦН	РАБ.	52847	1845.78	3238.08
4	d3b63715faebec4d81754c3e7a3ad836	ЭЦН	РАБ.	77725	2539.15	2937.34
5	5d995ad52dbe6eff8f1b6681d6b21bec	ЭЦН	РАБ.	58760	758.38	1422.56
6	ef3f72b14c3febb4d177c5e92ae4b561	ЭЦН	РАБ.	122280	2332.96	2271.86
7	429ecc93c8d72b9645bfd48c0f826dc7	ЭЦН	РАБ.	139720	1581.82	2548.04
8	5756c913481872253c810e989328d7a5	ЭЦН	РАБ.	132117	1944	2816
9	475a517d2c9ac61bd91a386b24824610	ЭЦН	РАБ.	148053	1706.47	2837.51
10	fe017a547ac58cc7366a60cc6bcd3915	ЭЦН	РАБ.	86073	864.8	1077.27

Рисунок 19. Выбор целевой функции

В таблице 3 представлены варианты, как в процентном соотношении разбивать исследуемый набор данных на тестовую и обучающую выборки. В данном случае процентное соотношение будет 70 % на 30% (70 % - тестовая выборка, 30 % - обучающая выборка).

Таблица 3. Выбор целевой функции, обучающейся и тестовой выборок

Objective function/ Целевая функция	**Binary** (0,1) Бинарная	
Training samples/ Обучающая выборка	**Sampling** 70%-80% Выборка	Representative relative to the objective function (GB)/ Репрезентативная по GB
Testing samples/ Тестовая выборка	**Sampling** 30%-20% Выборка	Representative relative to the objective function (GB)/ Репрезентативная по GB

3.2. Создание и тренировка линейного и полиномиального регрессоров

Первый шаг - разделение набора данных на тестовую и обучающую выборку. В первом случае создаётся линейный регрессор. С помощью данного регрессора строится предсказание количества добываемой нефти (в $м^3$) в зависимости от дебита жидкости. На рисунке 20 продемонстрирован результат работы.

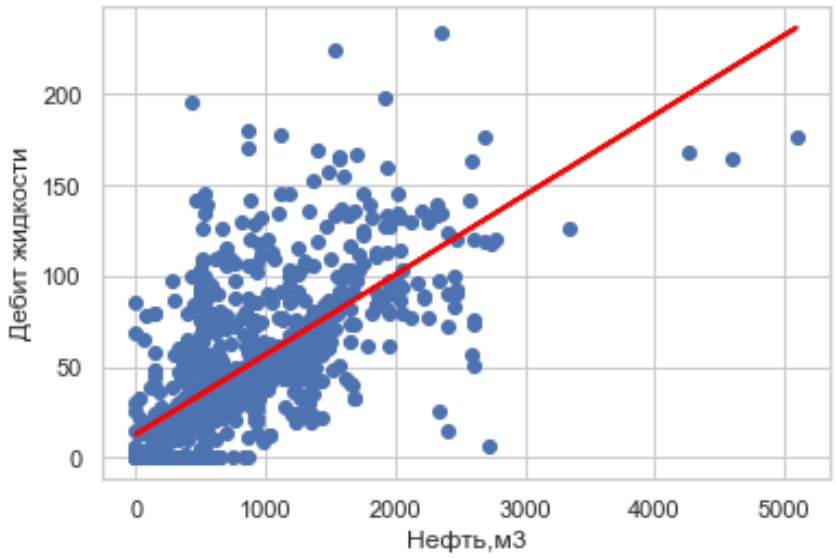


Рисунок 20. Линейная регрессия для одного параметра

В втором случае создаётся полиномиальный регрессор. С помощью данного регрессора строится предсказание количества добываемой нефти (в м3) в зависимости от дебита жидкости. На рисунке 21 продемонстрирован результат работы.

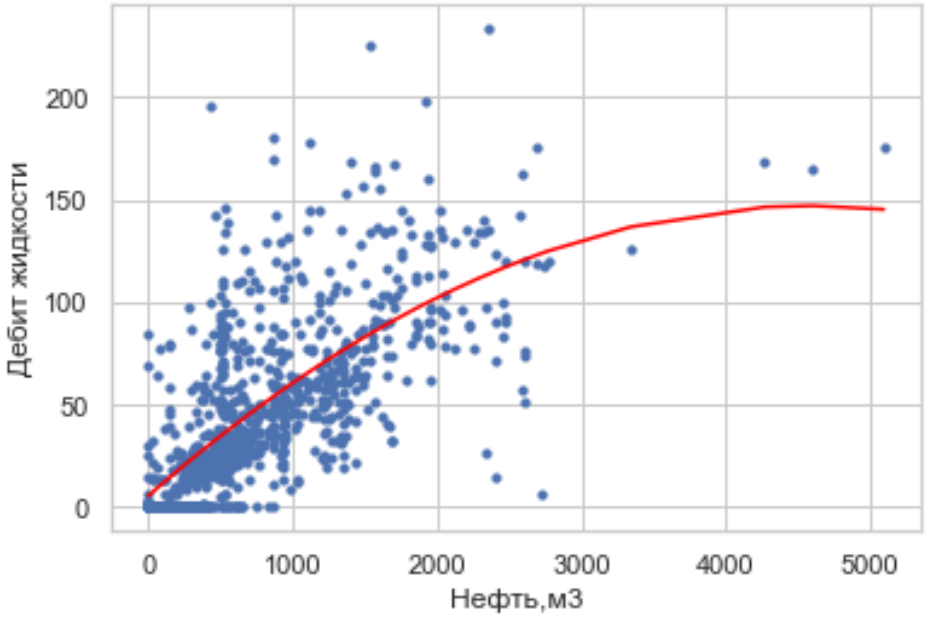


Рисунок 21. Полиномиальная регрессия для одного параметра

Для определения качества построенных моделей в данном исследовании будут использоваться метрики R2 и MSE (mean square error).

- *Mean Squared Error – среднеквадратичная ошибка.*

Измеряет среднюю сумму квадратной разности между фактическим значением и прогнозируемым значением для всех точек данных. Выполняется возведение во вторую степень, поэтому отрицательные значения не компенсируют положительными. А также в силу свойств этой метрики, усиливается влияние ошибок, по квадратуре от исходного значения. Чем меньше MSE, тем точнее предсказание.

- *R2 – коэффициент детерминации.*

Характеризует степень сходства исходных данных и предсказанных. В отличии от MSE не зависит от единиц измерения данных. R^2 показывает, насколько хорошо термины (точки данных) соответствуют кривой или линии. Если вы с каждой итерацией добавляете больше бесполезных переменных для модели, метрика R2 будет уменьшаться.

В таблице 4 представлены результаты оценки использованных регрессоров на тренировочном и тестовом образце набора данных.

Таблица 4. Сравнительный анализ регрессоров

Метрики	Линейная регрессия		Полиномиальная регрессия	
	Тренировочная выборка	Тестовая выборка	Тренировочная выборка	Тестовая выборка
MSE	8.3	15.36	7.9	11.32
R^2	88.8 %	81.4 %	90.1 %	84.3 %

На основании полученных результатов можно сделать вывод, что с увеличением степени полинома соответственно увеличивается процент точности построения регрессора. Стоит отметить, что при постоянном увеличении степени полинома точность построения регрессора может уменьшаться.

3.3. Построение и тренировка регрессора методом случайного леса

Случайный лес - модель, состоящая из множества деревьев решений. Вместо того, чтобы просто усреднять прогнозы разных деревьев (такая концепция называется просто «лес»), эта модель использует две ключевые концепции, которые и делают этот лес случайным:

1. *Случайная выборка образцов из набора данных при построении деревьев.*

В процессе тренировки каждое дерево случайного леса учится на случайном образце из набора данных. Выборка образцов происходит с возмещением (в статистике этот метод называется бутстреппинг, bootstrapping). Это даёт возможность повторно использовать образцы одним и тем же деревом. Хотя каждое дерево может быть высоковариативным по отношению к определённому набору тренировочных данных, обучение деревьев на разных наборах образцов позволяет понизить общую вариативность леса, не жертвуя точностью.

2. *При разделении узлов выбираются случайные наборы параметров.*

Вторая базовая концепция случайного леса заключается в использовании определённой выборки параметров образца для разделения каждого узла в каждом отдельном дереве. Обычно размер выборки равен квадратному корню из общего числа параметров.

Первый шаг - разделение набора данных на тестовую и обучающую выборку. С помощью данного регрессора методом случайного леса строится предсказание количества добываемой нефти (в м3) в зависимости от следующих параметров: добыча растворенного газа, жидкость, дебит попутного газа, диаметр эксплуатационной колонны, диаметр насосно-компрессорных труб (НКТ), глубина верхних дыр перфорации, производительность ЭЦН, глубина спуска, удельный коэффициент, коэффициент продуктивности.

В таблице 5 представлены результаты сравнения линейного, полиномиального регрессора с регрессором, полученным методом случайного леса.

Таблица 5. Сравнительный анализ регрессоров

Метрики	Лин. регрессия		Полином. регрессия		Случайный лес	
	Тренир. выборка	Тест. выборка	Тренир. выборка	Тест. выборка	Тренир. выборка	Тест. выборка
MSE	8.3	15.36	7.9	11.32	7	10.1
R^2	88.8 %	81.4 %	90.1 %	83.3 %	92 %	84.2 %

На основании полученных результатов можно сделать вывод, что регрессор, полученный методом случайного леса, имеет наибольшую точность для данного набора данных.

Стоит отметить, что в библиотеке для построения регрессора методом случайного леса есть функция, которая поможет определить наиболее важные признаки для выбранной целевой функции.

Это необходимо для того, чтобы исключить атрибуты, которые не несут никакой существенной информации для проведения дальнейших исследований.

В таблице 6 приведены самые важные признаки, расположенные в процентном соотношении.

Таблица 6. Важность признаков модели (в %)

Жидкость	30 %
Попутный газ	15 %
Дебит попутного газа	10 %
Добыча растворенного газа	10 %
Диаметр НКТ	5 %
Глубина верхних дыр перфорации	5 %
Производительность ЭЦН	4 %
Коэффициент продуктивности	3 %
Удельный коэффициент	2 %

Таким образом, исходя из полученных результатов можно сделать вывод о том, что на самом деле всего лишь 10 атрибутов из 43 имеют значительное влияние на предсказание о количестве добываемой нефти.

3.4. Вывод по разделу

В результате проведения регрессионного анализа были решены следующие задачи:

- Построение и тренировка линейного регрессора;
- Построение и тренировка полиномиального регрессора;
- Построение и тренировка регрессора, полученного методом случайного леса;
- Оценка точности каждой модели по определенным метрикам;
- Определить наиболее важные признаки для выбранной целевой функции.

В результате проведения разведочного и регрессионного анализа можно сделать вывод о том, что на предсказание о количестве добываемой нефти в рассматриваемом исследовании значительно влияют лишь 10 параметров (в начале исследования их было порядка 100 единиц).

4. Финансовый менеджмент, ресурсоэффективность и ресурсосбережение

4.1. Предпроектный анализ

4.1.1. Потенциальные потребители разработки

Диссертация посвящена разработке методологии обработки исходных данных с нефтегазового месторождения для построения модели машинного обучения. В данных условиях генерируются и обрабатываются сверхбольшие объемы данных, требующие иных подходов их обработки и хранения. Задача обработки больших данных остается актуальной и по сей день.

Нефтегазовые компании в процессе своей деятельности получают петабайты данных каждый день, использование больших данных открывает возможности анализа и предсказания развития трендов в области геологии, инженерии, производства и наилучшего способа использования оборудования для достижения наиболее оптимальных результатов работы на всех стадиях своей деятельности. Поэтому потенциальными потребителями, в первую очередь, являются нефтегазовые компании.

Актуальность данного раздела заключается в важности понимания коммерческой составляющей научно-технических проектов и оценки коммерческой ценности разработки.

Целью данного раздела является анализ совокупности факторов, которые определяют коммерческую привлекательность разработки, ее перспективность и успешность.

Основными задачами является оценка перспективности разработки, ее готовности к коммерциализации, выявление потенциальных угроз, а также расчет стоимости и составление графика проведения работ.

4.1.2. Технология QuaD

Технология QuaD позволяет оценить перспективность разработки на рынке и целесообразность вложения средств в научно-исследовательский проект. Результаты оценки, проведенной в табличной форме, представлены в таблице 7.

Таблица 7. QuaD-анализ разработки

Критерии оценки	Вес критерия	Средний балл	Макси-мальный балл	Относительное значение (3/4)	Средневзвешенное значение (5x2)
1	2	3	4	5	6
Производительность	0,07	80	100	0,8	0,04
Отказоустойчивость	0,17	90	100	0,9	0,153
Унифицированность	0,1	70	100	0,7	0,07
Безопасность	0,05	80	100	0,8	0,04
Потребность в ресурсах памяти	0,13	95	100	0,95	0,1235
Функциональная мощность	0,1	75	100	0,75	0,075
Простота эксплуатации	0,03	40	100	0,4	0,012
Масштабируемость	0,06	75	100	0,75	0,045
Конкурентоспособность продукта	0,07	50	100	0,5	0,035
Перспективность рынка	0,07	85	100	0,85	0,0595
Цена	0,1	40	100	0,4	0,04
Финансовая эффективность научной разработки	0,07	80	100	0,8	0,056
Итого	1				**0,749**

По результатам оценки качества и перспективности можно утверждать, что перспективность текущей разработки выше среднего. Улучшить данную разработку можно путем повышения качества пользовательского интерфейса.

4.1.3. SWOT-анализ

SWOT – Strengths (сильные стороны), Weaknesses (слабые стороны), Opportunities (возможности) и Threats (угрозы) – представляет собой комплексный анализ научно-исследовательского проекта. SWOT-анализ применяют для исследования внешней и внутренней среды проекта.

Разработанная для данного исследования матрица SWOT представлена в таблице ниже.

Таблица 8. Матрица SWOT разработки

	Сильные стороны научно-исследовательского проекта: С1. Не требуется специализированного оборудования С2. Невысокие системные требования	**Слабые стороны научно-исследовательского проекта:** Сл1. Ограниченный функционал конечного ПО. Сл2. Малый опыт создания подобных систем.
Возможности: В1. Невысокий уровень конкуренции В2. Разработка применима для различных предприятий	Возможность захватить рынок в разных сферах до появления конкурентов.	Благодаря невысокому уровню конкуренции, расширение функционала не будет приоритетной задачей.
Угрозы: У1. Низкий спрос у потребителя У2. Появление конкурентов в данном виде услуг	Низкие требования могут привлечь клиентов. В случае появления конкурентов в качестве преимущества можно рассматривать опыт в данной работе и сформировать базу клиентов.	Регулярная работа над проектом позволит найти новых клиентов, получить опыт и произвести новый функционал.

Данная разработка обладает рядом возможностей в условиях низкой вероятности возникновения угроз. Разработка спроектирована таких образом, что сильные стороны предусматривают изменение требований к самой методологии, а также возникновению задач по масштабированию разработки.

4.1.4. Оценка готовности разработки к коммерциализации

Одной из важных задач в ходе выполнения данного раздела является оценка готовности разработки к коммерциализации. Оцениваемыми параметрами являются как научная, так и коммерческая составляющая. Таблица 8 представляет собой бланк оценки степени готовности разработки к коммерциализации.

Таблица 8. Бланк оценки степени готовности разработки к коммерциализации

№ п/п	Наименование	Степень проработанности разработки	Уровень имеющихся знаний у разработчика
1.	Определен имеющийся научно-технический задел	4	4
2.	Определены перспективные направления коммерциализации научно-технического задела	4	5
3.	Определены отрасли и технологии (товары, услуги) для предложения на рынке	2	2
4.	Определена товарная форма научно-технического задела для представления на рынок	2	2
5.	Определены авторы и осуществлена охрана их прав	3	3
6.	Проведена оценка стоимости интеллектуальной собственности	3	3
7.	Проведены маркетинговые исследования рынков сбыта	1	1

8.	Разработан бизнес-план коммерциализации научной разработки	1	1
9.	Определены пути продвижения научной разработки на рынок	3	4
10.	Разработана стратегия (форма) реализации научной разработки	5	5
11.	Проработаны вопросы международного сотрудничества и выхода на зарубежный рынок	5	4
12.	Проработаны вопросы использования услуг инфраструктуры поддержки, получения льгот	2	2
13.	Проработаны вопросы финансирования коммерциализации научной разработки	2	3
14.	Имеется команда для коммерциализации научной разработки	3	3
15.	Проработан механизм реализации разработки	5	5
	ИТОГО БАЛЛОВ:	45/75	47/75

Поскольку данная разработка является индивидуальным проектом для уникального научного проекта, не предполагающем дальнейший выход на рынок, коммерциализация данного продукта не является целесообразной. В связи с этим провести полноценную оценку перспективы коммерциализации не представляется возможным. По результатам оценки можно утверждать, что данный проект еще не готов к коммерциализации, главным образом, с точки зрения сбыта разработки и финансирования коммерциализации.

4.2. Инициация разработки

В рамках инициации разработки формулируются цели и ожидаемые результаты работы. Также определяются заинтересованные стороны разработки и возможные ограничения. Заинтересованные в данной разработке стороны представлены в таблице 9.

Таблица 9. Заинтересованные стороны разработки

Заинтересованные стороны	**Ожидания заинтересованных сторон**
Нефтегазовые компании	Сокращение затрат и увеличение эффективности работы оборудования и персонала, увеличение прибыли
Разработчики в области обработки больших данных	Получения опыта и создание новых и эффективных методологий для обработки больших объемов данных

Цели и результат проекта отображены в таблице 10.

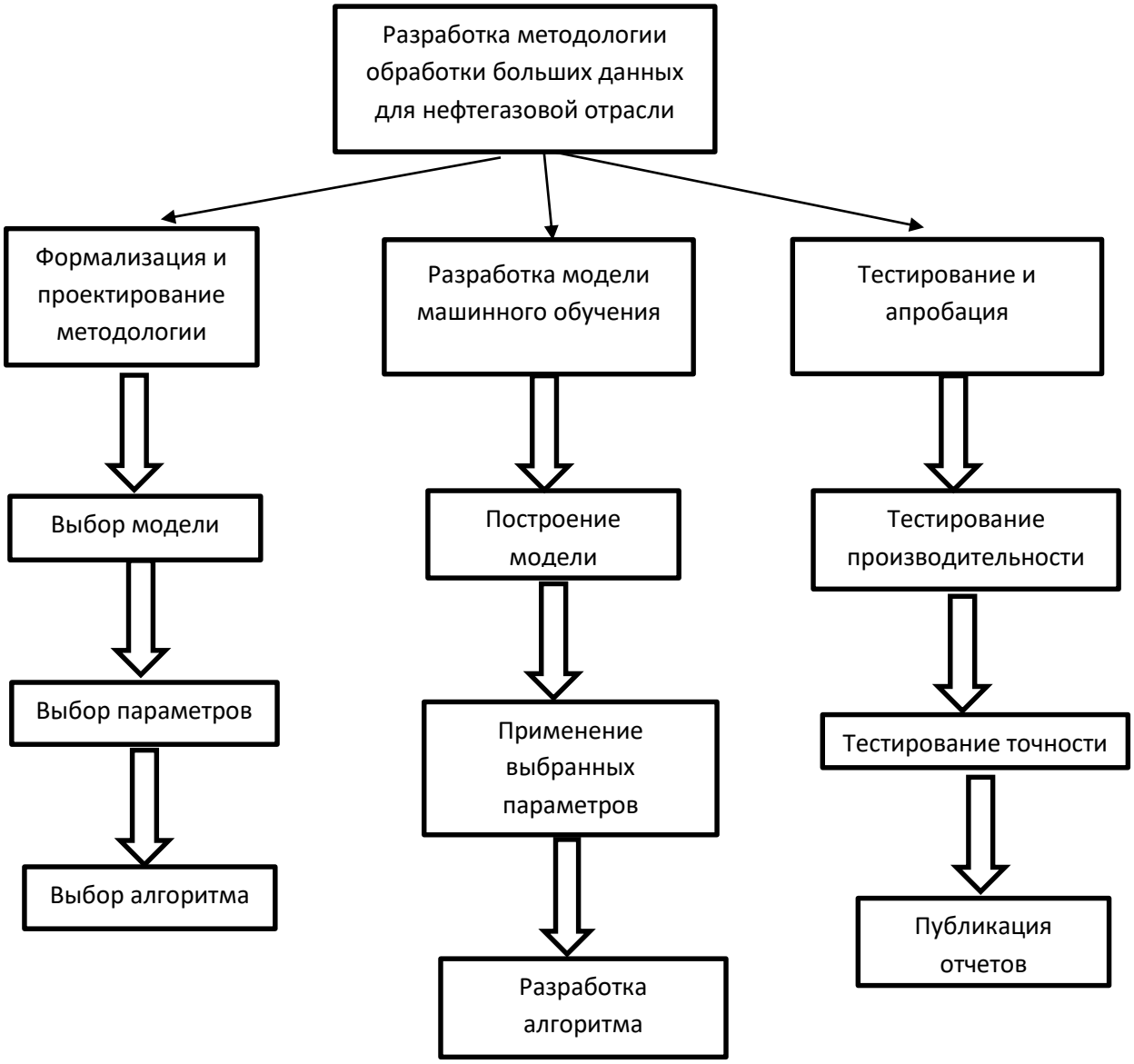
Таблица 10. Цели и результат разработки

Цели разработки:	**Разработка методологии обработки больших данных для нефтегазовой отрасли**
Ожидаемые результаты разработки:	1) Формализованное описание методологии 2) Программная реализация алгоритмов
Критерии приемки результата разработки:	1) Точность работы алгоритмов 2) Эффективность работы алгоритмов
Требования к результату разработки:	**Требования:**
	Формализованное описание работы методологии
	Точность работы алгоритмов

4.3. Планирование управления разработкой

4.3.1. Иерархическая структура работ

Иерархическая структура работ для данной разработки представляет собой детализацию укрупненной структуры работ, продемонстрированной ниже.



Задачи по созданию данной разработки разделены на три основных блока: формализация и разработка, реализация, а также тестирование и апробация.

4.3.2. План разработки

Чтобы отразить ключевые события по ведению разработки, необходимо составить календарный план.

Таблица 11. Календарный план разработки

Код работы	Название	Длительность, дни	Дата начала работ	Дата окончания работ	Состав участников
1	Выбор научного руководителя магистерской работы	1	01.09.19	01.09.19	Журбич Никита Игоревич
2	Составление и утверждение темы магистерской работы	2	03.09.19	04.09.19	Губин Евгений Иванович
3	Составление календарного плана-графика выполнения магистерской работы	2	05.09.19	06.09.19	Журбич Никита Игоревич, Губин Евгений Иванович
4	Выявление требований к разработке	7	07.09.19	14.09.19	Журбич Никита Игоревич, Губин Евгений Иванович
5	Подбор и изучение литературы по теме магистерской работы	25	15.09.19	13.10.19	Журбич Никита Игоревич
6	Анализ предметной области	15	15.10.19	31.10.19	Журбич Никита Игоревич
7	Формализация протокола передачи данных	30	01.11.19	05.12.19	Журбич Никита Игоревич

8	Разработка протокола	80	06.12.19	08.03.20	Журбич Никита Игоревич
9	Тестирование	20	09.03.20	01.04.20	Журбич Никита Игоревич
10	Анализ полученных результатов, сравнительная оценка производительности протокола	4	02.04.20	05.04.20	Журбич Никита Игоревич, Губин Евгений Иванович
11	Согласование выполненной работы с научным руководителем	4	06.04.20	10.04.20	Журбич Никита Игоревич, Губин Евгений Иванович

4.3.2.1. Продолжительность этапов работ

Трудоемкость выполнения научного исследования оценивается экспертным путем в человеко-днях и носит вероятностный характер, завися от множества трудно учитываемых факторов. Для определения ожидаемого значения трудоемкости $t_{ожi}$ используется формула:

$$t_{ожi} = \frac{3tmini+2tmaxi}{5}, \qquad (2)$$

где $t_{ожi}$ – ожидаемая трудоемкость i-й работы чел.-дн;

t_{maxi} – минимально возможная трудоемкость выполнения заданной работы, чел.-дн.;

t_{mini} – минимально возможная трудоемкость выполнения заданной работы, чел.-дн.

Промежуточные расчеты представлены в таблице 12.

Таблица 12. Временные показатели проведения разработки

Наименование работы	Исполнители работы	Трудоемкость работ, чел-дни			Длительность работ, дни	
		tmin	tmax	tож	Тр	Тк
Выбор научного руководителя магистерской работы	Журбич Н.И.	1	2	1,4	1	1
Составление и утверждение темы магистерской работы	Губин Е.И.	1	3	1,8	2	2
Составление календарного плана-графика выполнения магистерской работы	Журбич Н.И.	1	3	1,8	2	2
	Губин Е.И.	1	3	1,8	2	2
Выявление требований к разработке	Журбич Н.И.	5	10	7	7	8
	Губин Е.И.	5	10	7	7	8
Подбор и изучение литературы по теме магистерской работы	Журбич Н.И.	22	30	25,2	25	29
Анализ предметной области	Журбич Н.И.	10	22	14,8	15	17
Формализация протокола передачи данных	Журбич Н.И.	20	45	30	30	35
Разработка протокола	Журбич Н.И.	60	110	80	80	89
Тестирование	Журбич Н.И.	10	35	20	20	24
Анализ полученных результатов, сравнительная оценка производительности протокола	Журбич Н.И.	2	7	4	4	4
	Губин Е.И.	2	7	4	4	4
Согласование выполненной работы с научным руководителем	Журбич Н.И.	2	7	4	4	5
	Губин Е.И.	2	7	4	4	5
Выполнение других частей работы	Журбич Н.И.	20	45	30	30	35
Подведение итогов, оформление работы	Журбич Н.И.	7	14	9,8	10	12
	Губин Е.И.	7	14	9,8	10	12

4.3.2.2. Разработка графика проведения разработки

На рисунке представлена диаграмма Ганта с планом выполнения работ, где М – магистрант (Журбич Н.И.), НР – научный руководитель (Губин Е.И.).

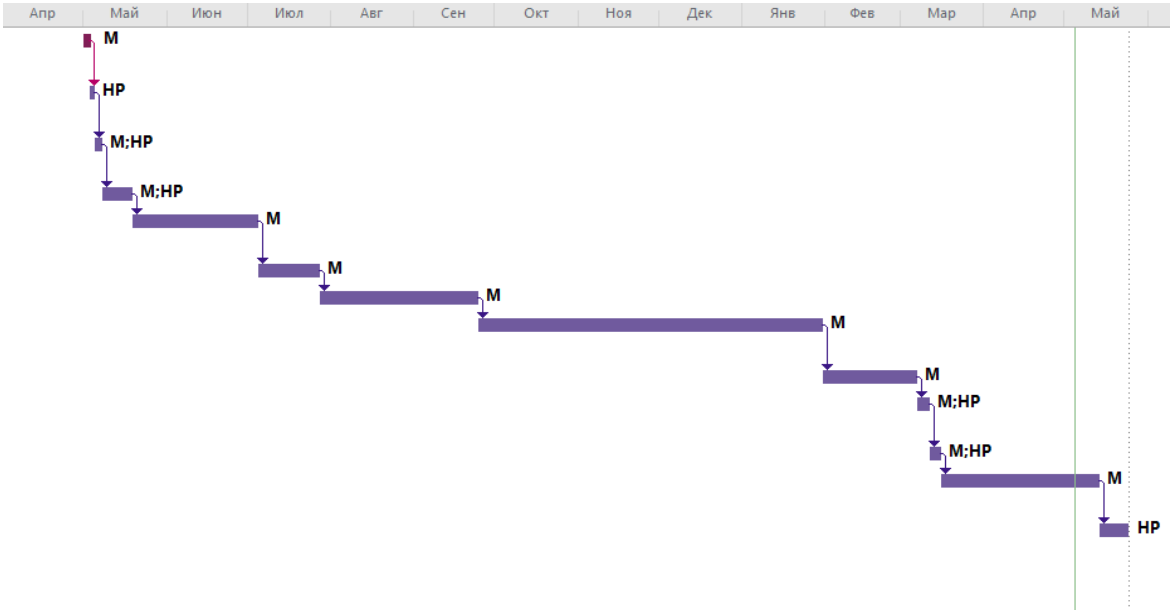


Рисунок 22. График проведения разработки

4.3.3. Бюджет разработки

Для данной разработки бюджет состоит из следующих пунктов:

1) Материальные затраты;

2) Амортизационные отчисления;

3) Основная заработная плата исполнителей темы;

4) Дополнительная заработная плата исполнителей темы;

5) Страховые отчисления;

6) Накладные расходы.

4.3.3.1. Расчет материальных затрат разработки

Для проведения исследования какие-либо специальные материалы и комплектующие не приобретались. Единая сумма на канцелярские принадлежности составляет 2500 рублей.

4.3.3.2. Расчет амортизационных отчислений

Поскольку для проведения исследований специальное дорогостоящее оборудование не приобреталось, при расчете затрат учитывается только амортизация. Первоначальная стоимость ПК магистранта, используемого для проведения исследований, составляет 50000 рублей. Срок полезного использования данной машины – 3 года, из которых 9 месяцев машина использовалась для написания ВКР.

Норма амортизации:

$$А_н = \frac{1}{n} \times 100\% = \frac{1}{3} \times 100\% = 33.33\ \% \qquad (3)$$

Годовые амортизационные отчисления:

$$А_г = 50000 \times 0{,}33 = 16500 \text{ рублей} \qquad (4)$$

Ежемесячные амортизационные отчисления:

$$А_м = \frac{16500}{9} = 1833{,}3 \text{ рублей} \qquad (5)$$

Итоговая сумма амортизации основных средств:

$$А = 1833{,}3 \times 9 = 16\ 499 \text{ рублей} \qquad (6)$$

Таким образом затраты на амортизацию ПК составляют 16 499 рублей.

Суммарная стоимость остального программного обеспечения с перманентной лицензией составляет 42000 рублей (в стоимость также включена годовая лицензия программного среды разработки для Python).

Норма амортизации:

$$А_н = \frac{1}{n} \times 100\ \% = \frac{1}{5} \times 100\ \% = 20\ \% \qquad (7)$$

Годовые амортизационные отчисления:

$$А_г = 42000 \times 0{,}2 = 8400 \text{ рублей} \qquad (8)$$

Ежемесячные амортизационные отчисления:

$$А_м = 8400 / 9 = 933 \text{ рублей} \quad (9)$$

Итоговая сумма амортизации основных средств:

$$А = 933 \times 9 = 8397 \text{ рублей} \quad (10)$$

Амортизация остального программного обеспечения составляет 8397 рублей. Итого амортизация ПО составляет 25 436 рублей.

Таблица 13. Расчет затрат на амортизацию

Наименование	Затраты, руб.
Амортизация ПК	16 499
Амортизация ПО	25 436
Итого	**41 935**

4.3.3. Основная заработная плата исполнителей темы

Исполнителями темы выступают научный руководитель и инженер. Оклад руководителя в ТПУ без районного коэффициента составляет 33664 рубля, оклад инженера – 21760 рублей. Баланс рабочего времени для 6-дневной недели, по которой учитывается рабочее время преподавателей и студентов, представлен в таблице.

Таблица 14. Расчет основной заработной платы

Исполнители	Здн, руб.	Кпр	Кд	Кр	Тр	Зосн
Инженер	931,29	0,3	0,2	1,3	228	414051,53
Руководитель	1440,76	0,3	0,5	1,3	29	97769,97
					Итого:	**511821,5**

$$З_{дн} = \frac{Зм \times М}{Fд} = \frac{21760 \times (1+Кпр+Кд) \times Кр \times 10,4}{243} = 1860,02 \text{ руб.} \quad (11)$$

$$З_{дн} = \frac{Зм \times М}{Fд} = \frac{33664 \times (1+0,3+0,5) \times 1,3 \times 10,4}{243} = 3371,38 \text{ руб.} \quad (12)$$

Расчет основной заработной платы (для инженера и руководителя соответственно):

$$З_{осн} = З_{дн} \times Т_р \quad (13)$$

$$Зосн_{инж} = 1860,02 \times 228 = 414051,53 \quad (14)$$

$$Зосн_{рук} = 3371,38 \times 29 = 97769,97 \quad (15)$$

4.3.3.4. Дополнительная заработная плата исполнителей темы

Пусть дополнительная заработная плата составляет 15% от основной заработной платы. Тогда зарплаты инженера и руководителя соответственно будут высчитываться по формуле:

$$З_{доп} = З_{осн} \times 0,15 \quad (15)$$

$$Здоп_{инж} = 414051,53 \times 0,15 = 62107,73 \quad (16)$$

$$Здоп_{рук} = 97769,97 \times 0,15 = 14665,5 \quad (17)$$

Дополнительная заработная плата исполнителей равна 62107,73 и 14665,5 рублей соответственно.

4.3.3.5. Отчисления во внебюджетные фонды (страховые отчисления)

Составляют 30% от заработной платы (основная + дополнительная). Таким образом страховые взносы составляют 176578,42 рублей.

$$Отч = (З_{доп} + З_{осн}) \times 0,3 \quad (18)$$

$$Отч = (511\,821,5 + 76\,773,23) \times 0,3 = 176578,42 \quad (19)$$

4.3.3.6. Накладные расходы

Накладные расходы составляют 16% от суммы материальных затрат, затрат на специальное оборудование, затрат на основную заработную плату, затрат на дополнительную заработную плату и страховых взносов. Накладные расходы составляют 128500,664 рублей.

4.3.3.7. Формирование бюджета затрат научно-исследовательского разработки

Таблица 15. Бюджет затрат

Наименование	Сумма, руб.	Удельный вес, %
Материальные затраты	2500	0,27
Затраты на специальное оборудование	41 935	4,06
Затраты на основную заработную плату	511821,5	54,77
Затраты на дополнительную заработную плату	76773,23	8,22
Отчисления во внебюджетные фонды	176578,42	18,89
Накладные расходы	128891,06	13,79
Общий бюджет	938 499,21	100

Общий бюджет разработки составляет 938 499,21 рублей. Данный бюджет ниже конкурентных зарубежных разработок.

4.3.4. Риски разработки

Проведение любого научно-исследовательского проекта сопряжено с возникновением различных рисков. Предварительное определение рисков помогает своевременному принятию мер по предотвращению возникновения угроз или минимизации их последствий.

Таблица 16. Определение рисков

№	Наименование риска	Описание риска
1	Политические	Риск отказа от продолжения сотрудничества потенциального заказчика в связи с обострением политической обстановки.
2	Технологические	Безвозвратная утеря большого процента исходных данных, на которые опирается разработка, в ходе работы.
3	Финансовые	Прекращение финансирования проекта.
4	Технические	Сбой или поломка оборудования, связанного с хранилищами данных, на которые опирается разработка.

Таблица 17. Оценка вероятности рисков

№ п/п	Наименование риска	Оценка вероятности риска (низкая, средняя, высокая)
1	Политические	Низкая
2	Технологические	Низкая
3	Финансовые	Низкая
4	Технические	Низкая

Таблица 18. Оценка уровня потерь

№ п/п	Наименование риска	Оценка уровня потерь (низкий, средний, высокий)
1	Политические	Высокий
2	Технологические	Средний
3	Финансовые	Высокий
4	Технические	Низкий

Таблица 19. Основные мероприятия по снижению рисков

№ п/п	Наименование риска	Мероприятия по снижению риска
1	Политические	Заключение контракта о сотрудничестве на четко обозначенный период.
2	Технологические	Разграничение уровня доступа пользователей. Создание бэкапов данных.
3	Финансовые	Своевременное принятие мер по подготовке отчетности по текущим работам и подача заявки на новые.
4	Технические	Соблюдение протокола безопасности

Основными рисками при выполнении разработки можно назвать технологические и технические риски, связанные с хранилищами данных о научном эксперименте в области обработки больших объемов данных, при этом наибольшую угрозу представляют собой технологические риски. Среди прочих рисков стоит отметить, что финансовые риски связаны, в основном, с финансированием. Предотвращения данного риска возможно в случае своевременного предоставление релевантной документации в научные фонды. Политические риски наименее вероятны, однако стоит отметить, что в случае наступления предполагаемого сценария, подобные риски имеют серьезные последствия для такого уникального научного проекта.

4.4. Определение потенциального эффекта разработки

Потенциальным пользователям разработки являются нефтегазовые компании, поэтому метод оценки абсолютной эффективности исследования не подходит для данного исследования. Поскольку в ходе выполнения магистерской диссертации разрабатывался только один вариант разработки, провести оценку сравнительной эффективности исследования не представляется возможным.

Данная разработка ориентирована на конкретного потребителя/группу потребителей, которыми являются нефтегазовые компании. Стоит отметить, что, поскольку потенциальный рынок сбыта мал, коммерциализация данной разработки остается открытым вопросом. Бланк оценки степени готовности к коммерциализации указывает на то, что готовность в коммерциализации находится на нижнем пороге уровня выше среднего.

Потенциальная стоимость исследования составляет около 1 млн. рублей. При этом специальное дорогостоящее оборудование не закупалось. Данная цена является конкурентоспособной, поскольку стоимость решений для подобных вычислительных структур «из коробки», не ориентированных на конкретного потребителя, находятся в том же ценовом диапазоне.

Данный проект сопряжен с малыми рисками, однако стоит обратить внимание на технологические риски, последствия которых могут нанести существенный урон подобным исследованиям, а также при продолжении работы над данным проектом. Проект не имеет прямых аналогов в обозначенных условиях обработки больших объемов данных в нефтегазовой области.

4.5. Выводы по разделу

Проведено комплексное описание и анализ финансово-экономических аспектов выполненной работы.

Составлен перечень проводимых работ, их исполнителей и продолжительность выполнения этапов работ, составлен линейный график.

Рассчитана смета затрат на выполнение проекта, проведен расчет себестоимости и прибыли проекта.

Определены показатели эффективности проекта и проведена оценка его эффективности.

5. Социальная ответственность

В данной главе освещен комплекс мер организационного, правового, технического и режимного характера, которые минимизируют негативные последствия разработки программного комплекса, а также рассматриваются вопросы техники безопасности, охраны окружающей среды и пожарной профилактики, даются рекомендации по созданию оптимальных условий труда.

Объектом исследования выступает рабочее место программиста, разрабатывающего данную методологию, которая позволить корректно обрабатывать исходные данные с нефтегазового месторождения.

Рабочей зоной при разработке данной методологии является учебная аудитория в Кибернетическом центре ТПУ, оборудованная системой отопления, кондиционирования воздуха, с естественным и искусственным освещением.

Рабочее место – стационарное, оборудованное персональным компьютером и оргтехникой.

5.1. Правовые и организационные вопросы обеспечения безопасности

5.1.1. Специальные (характерные для проектируемой рабочей зоны) правовые норма трудового законодательства.

Правовое регулирование трудовых отношений между работодателем, работником и государством регулируется Трудовым кодексом Российской Федерации от 30.12.2001 N 197-ФЗ. В ТК РФ [9], в соответствии с Конституцией РФ, признаются свобода труда, выбор и согласие на него, а также выбор профессии и деятельности. Запрещаются принудительный труд, дискриминация по какому-либо признаку. Гарантируются справедливые и достойные условия труда.

ТК РФ регламентирует порядок разрешения индивидуальных и коллективных трудовых споров, особенности труда женщин, детей и людей пенсионного возраста, права и обязанности работодателей и работников,

нормы рабочего времени, порядок оплаты труда и виды компенсаций во вредных условиях труда, а также особенности социального страхования.

В соответствии со ст. 111 ТК РФ, рабочая неделя (в т.ч. шестидневная) не должна превышать 40 часов в неделю. Воскресенье является выходным днем.

В соответствии со ст. 212 ТК РФ, работодатель обязан обеспечить безопасные условия труда, а также обязательное социальное страхование работников от несчастных случаев на производстве и профессиональных заболеваний.

В соответствии со ст. 142 ТК РФ, в случае задержки выплаты заработной платы на срок более 15 дней работник имеет право, известив работодателя в письменной форме, приостановить работу на весь период до выплаты задержанной суммы, кроме ряда перечисленных случаев.

Данные о работнике, предоставляемые работодателю, обрабатываются только с согласия самого работника и охраняются Федеральным Законом от 27.07.2006 N 152-ФЗ (ред. от 25.07.2011) «О Персональных Данных» [10].

5.1.2. Организационные мероприятия при компоновке рабочей зоны

Согласно ГОСТ 12.2.032-78 ССБТ [11], выявлены следующие параметры рабочей зоны:

- Согласно наименованию работы (работа за ЭВМ) при отсутствии регулирующих механизмов высоты рабочей поверхности, высота рабочей поверхности, при организации рабочего места, составляет (для мужчин) 700 мм. Высота сиденья 500 мм.
- Рабочая поверхность в соответствии с видом работ может содержать дополнительное углубление для периферийных устройств (клавиатура).
- Рабочее место при выполнении работ сидя организуют при легкой работе, не требующей свободного передвижения работающего.

- Конструкция рабочего места и взаимное расположение всех его элементов должны соответствовать антропометрическим, физиологическим и психологическим требованиям, а также характеру работы.

В соответствии с СанПиН 2.2.2/2.4.1340-03 «Гигиенические требования к персональным электронно-вычислительным машинам и организации работы» (с изменениями на 21 июня 2016 года) [12] были выявлены следующие правила организации работы с ПЭВМ:

- Освещенность на поверхности стола в зоне размещения рабочего документа должна быть 300-500 лк. Освещение не должно создавать бликов на поверхности экрана. Освещенность поверхности экрана не должна быть более 300 лк.
- При размещении рабочих мест с ПЭВМ расстояние между рабочими столами с видеомониторами (в направлении тыла поверхности одного видеомонитора и экрана другого видеомонитора), должно быть не менее 2,0 м, а расстояние между боковыми поверхностями видеомониторов - не менее 1,2 м.
- Конструкция рабочего стула (кресла) должна обеспечивать поддержание рациональной рабочей позы при работе на ПЭВМ, позволять изменять позу с целью снижения статического напряжения мышц шейно-плечевой области и спины для предупреждения развития утомления. Тип рабочего стула (кресла) следует выбирать с учетом роста пользователя, характера и продолжительности работы с ПЭВМ.
- Ширина и глубина поверхности сиденья должны составлять не менее 400 мм.
- Контролируемыми гигиеническими параметрами персональных цифровых ЭВМ (в т.ч. портативных) являются: уровни электромагнитных полей (ЭМП), акустического шума, концентрация вредных веществ в воздухе, визуальные показатели ВДТ, мягкое рентгеновское излучение.

5.2. Профессиональная социальная безопасность

В данном подразделе анализируются вредные и опасные факторы, которые могут возникать при проведении исследований в лаборатории, при разработке или эксплуатации проектируемого решения.

Согласно ГОСТ 12.0.003-2015 [13] в таблице 20 представлены возможные вредные и опасные факторы. Работа по разработке программного обеспечения делится на три основных этапа: проектирование, разработка и эксплуатация.

Таблица 20. Возможные вредные и опасные факторы

Факторы (ГОСТ 12.0.003-2015)	Этапы работ			Нормативные документы
	Проектирование	Разработка	Эксплуатация	
1. Отклонение показателей микроклимата	+	+	+	- СанПиН 2.2.4.548–96 - СанПиН 2.2.2/2.4.1340-03
2. Превышение уровня шума	+	+	+	- ГОСТ 12.1.003-2014 ССБТ - ГОСТ 12.1.029-80
3. Отсутствие или недостаток естественного света	+	+	+	- СанПиН 2.2.1/2.1.1.1278–03
4. Недостаточная освещенность рабочей зоны	+	+	+	- СанПиН 2.2.1/2.1.1.1278–03
5. Повышенное значение напряжения в электрической цепи	+	+	+	- ГОСТ 12.1.019-2017

5.2.1. Анализ вредных и опасных факторов, которые может создать объект исследования.

Теоретически, объект исследования (методология подготовки исходных данных) при определенных условиях, не предусмотренных разработчиками, способен привести к перегрузкам в ПЭВМ и вызвать определенные последствия. Поскольку разработка не является осязаемым объектом и неотделима от ПЭВМ, вредные и опасные факторы, которые могут быть прямо или косвенно отнесены к разработке, относятся и к рабочему месту.

5.2.2. Анализ вредных и опасных факторов, которые могут возникнуть на рабочем месте.

5.2.2.1. Отклонение показателей микроклимата.

Показатели микроклимата должны обеспечивать сохранение теплового баланса человека с окружающей средой и поддержание оптимального или допустимого теплового состояния организма. Отклонение показателей от нормы в пределах допустимых величин могут вызвать локальные ощущения теплового дискомфорта и напряжение механизмов терморегуляции.

Согласно СанПиН 2.2.4.548-96 [14], санитарные правила устанавливают гигиенические требования к показателям микроклимата рабочих мест производственных помещений с учетом интенсивности энерготрат работающих, времени выполнения работы, периодов года и содержат требования к методам измерения и контроля микроклиматических условий.

Оптимальные микроклиматические условия установлены по критериям оптимального теплового и функционального состояния человека. Они обеспечивают общее и локальное ощущение теплового комфорта в течение 8-часовой рабочей смены при минимальном напряжении механизмов терморегуляции, не вызывают отклонений в состоянии здоровья, создают

предпосылки для высокого уровня работоспособности и являются предпочтительными на рабочих местах.

Категория работ определена, как наименее энергозатратная, т.е. не сопровождающаяся какой-либо физической нагрузкой (сидячий вид деятельности, умственный труд) – 1а. Оптимальные величины показателей микроклимата представлены в таблице 21.

Таблица 21. Оптимальные величины показателей микроклимата на рабочих местах производственных помещений.

Период года	Категория работ по уровню энергозатрат, Вт	Температура воздуха, °С	Температура поверхностей, °С	Относительная влажность воздуха, %	Скорость движения воздуха, м/с
Теплый	1а	22-24	21-25	60-40	0,1
Холодный	1а	23-25	22-26	60-40	0,1

Для определения температуры воздуха в помещении использовался настенный термометр. Показания температуры воздуха составляют около 23 °С. Для измерения относительной влажности воздуха использовался гигрометр механического типа. Показания гигрометра на момент измерения составили 45%. Отсюда можно сделать вывод, что микроклимат в учебной аудитории является оптимальным для работы в данное время года.

5.2.2.2. Превышение уровня шума

Превышения уровня шума является вредным фактором на рабочем месте. Постоянный шум, превышающий допустимые значения, не только воздействует на органы слуха, но и влияет на общее самочувствие работника, способствует ослаблению организма, а также снижает работоспособность.

Согласно ГОСТ 12.1.003-2014 [15], машины, которые в процессе работы могут производить шум, неблагоприятно воздействующий на работников, следует конструировать и изготовлять с учетом последних

достижений технологии и принципов проектирования, позволяющих снизить излучаемый шум.

Наибольшим шумовым событием на рабочем месте для исследователя является ПЭВМ или несколько ПЭВМ.

СанПиН 2.2.2/2.4.1340-03 регулирует допустимые значения уровней звукового давления в октавных полосах частот и уровня звука, создаваемого ПЭВМ. Данные представлены в таблице 22.

Таблица 22. Допустимые значения уровней звукового давления в октавных полосах частот и уровня звука, создаваемого ПЭВМ.

Уровни звукового давления в октавных полосах со среднегеометрическими частотами									Уровни звука в дБА
31, 5 Гц	63 Гц	125 Гц	250 Гц	500 Гц	1000 Гц	2000 Гц	4000 Гц	8000 Гц	
86 дБ	71 дБ	61 дБ	54 дБ	49 дБ	45 дБ	42 дБ	40 дБ	38 дБ	50

5.2.2.3. Расчет искусственного освещения.

Основной задачей светотехнических расчётов для искусственного освещения является определение требуемой мощности электрической осветительной установки для создания заданной освещённости.

В расчётном задании должны быть решены следующие вопросы:

– выбор системы освещения;

– выбор источников света;

– выбор светильников и их размещение;

– выбор нормируемой освещённости;

– расчёт освещения методом светового потока.

Основные работы проводились в аудитории № 204 Кибернетического центра (корпус № 22). Поэтому было предложено рассчитать освещённость в данной аудитории.

- Выбор системы освещения

В данном расчётном задании для учебной аудитории рассчитывается общее равномерное освещение.

- Выбор источников света

Источники света, применяемые для искусственного освещения, делят на две группы – газоразрядные лампы и лампы накаливания. Для общего освещения, как правило, применяются газоразрядные лампы как энергетически более экономичные и обладающие большим сроком службы. Наиболее распространёнными являются люминесцентные лампы. Широко применяются люминесцентные лампы типа ЛБ. Именно такие установлены рассматриваемой аудитории.

- Выбор светильников и их размещение

При выборе типа светильников следует учитывать светотехнические требования, экономические показатели, условия среды.

Наиболее распространёнными типами светильников для люминесцентных ламп являются: открытые двухламповые светильники типа ОД, ОДОР, ШОД, ОДО, ООД – для нормальных помещений с хорошим отражением потолка и стен, допускаются при умеренной влажности и запылённости.

Аудитория имеет следующие параметры: длина A = 12 м, ширина B = 6 м, высота H = 4 м. Высота рабочей поверхности $h_{рп} = 0,8$ м.

Определим расчетную высоту подвеса светильников над рабочей поверхностью (h) по формуле:

$$h = H - h_c - h_p, \quad (20)$$

где H – высота помещения;

h_c – расстояние светильников от перекрытия (свес);

h_n = расстояние от пола до рабочей поверхности стола.

$$h = 4 - 0,5 - 0,2 = 3,3 \text{ м};$$

Индекс помещения определяется по формуле (21):

$$i = S/ h \times (A+B), \quad (21)$$

где S – площадь помещения, м^2;

A – длина комнаты, м;

B – ширина комнаты, м;

h – высота подвеса светильников, м.

$$i = 72 / [3, 3(12 +6)] = 1,21.$$

Исходя из того, что потолок в помещении чистый бетонный, а также свежепобеленные стены без окон, согласно методическим указаниям, примем коэффициенты отражения от стен ρ_c*=70%* и потолка ρ_n*=50%*. По таблице коэффициентов использования светового потока для соответствующих значений i, ρ_c, ρ_n, примем η=0,61.

Освещенность помещения рассчитывается по формуле (22):

$$E_ф = (N \times \eta \times Ф) / S \times K_З \times Z, \quad (22)$$

где $Ф$ – световой поток светильника, лм;

S – площадь помещения, м2;

$k_з$ – коэффициент неравномерности освещения;

n – число светильников;

η – коэффициент использования светового потока.

Коэффициент запаса k учитывает запыленность светильников и их износ. Для помещений с малым выделением пыли k = 1,5. Поправочный коэффициент z – это коэффициент неравномерности освещения. Для люминесцентных ламп z = 1,1. В помещении находятся светильники ЛВО

4×20 CSVT, с люминесцентными лампами типа L 20W/640 с потоком F = 1200 лм.

Учитывая все параметры, рассмотренные выше, найдем освещенность по формуле (22):

$$E_ф = \frac{60 \times 0{,}61 \times 1200}{72 \times 1{,}5 \times 1{,}1} = 369 \text{ лк.}$$

В рассматриваемом помещении освещенность должна составлять 300 лк согласно СНиП 23-05-95. В данном помещении освещенность превышает норму, следовательно, дополнительные источники света не нужны.

5.2.2.4. Умственное перенапряжение

Работая за ПЭВМ, работник также находится под влиянием еще одного вредного производственного фактора, нервно-психической перегрузки в виде умственного перенапряжения.

Согласно ТОИ Р-45-084-01 [16], виды трудовой деятельности разделяются на 3 группы: группа А - работа по считыванию информации с экрана компьютера с предварительным запросом; группа Б - работа по вводу информации; группа В - творческая работа в режиме диалога с компьютером. При выполнении в течение рабочей смены работ, относящихся к различным видам трудовой деятельности, за основную работу с компьютером следует принимать такую, которая занимает не менее 50% времени в течение рабочей смены или рабочего дня. Уровень нагрузки представлен в таблице 23.

Таблица 23. Уровень нагрузки за рабочую смену при видах работ с компьютером

Категория работ	Уровень нагрузки за рабочую смену при видах работ с компьютером		
	группа А, количество знаков	группа Б, количество знаков	группа В, час.
III	До 60000	До 40000	До 6,0

5.2.3. Обоснование мероприятий по защите исследователя от действия опасных и вредных факторов

Для поддержания оптимального микроклимата в помещении используются средства центрального отопления и кондиционирования в зависимости от сезона. Согласно СанПиН 2.2.2/2.4.1340-03, также:

1) В помещениях, оборудованных ПЭВМ, проводится ежедневная влажная уборка и систематическое проветривание после каждого часа работы на ЭВМ;
2) Уровни положительных и отрицательных аэроионов в воздухе помещений, где расположены ПЭВМ, должны соответствовать действующим санитарно-эпидемиологическим нормативам.

Для защиты от шума, согласно ГОСТ 12.1.029-80 ССБТ [17], могут применяться следующие средства и методы:

1) Рациональное размещение рабочих мест.
2) Рациональное размещение технологического оборудования.
3) Применение малошумных современных ПЭВМ.

Поскольку применение индивидуальных средств шумоизоляции оказывает дополнительный психологический эффект и ухудшение общего состояния работника после длительного использования, средства защиты от шума должны являться некоторым компромиссом, направленным на снижение уровня шума и, вместе с тем, сохранение комфортных условий работы.

Защитой от прямого прикосновения к токопроводящим частям электрооборудования являются:

1) Основная изоляция.
2) Безопасное расположение токоведущих частей.
3) Защитное отключение.

Окружающая среда не должна быть проводящей. При эксплуатации электрооборудования необходимо соблюдать технику безопасности.

Для III категории работ (ТОИ Р-45-084-01) по уровню нагрузки перерыв регламентирован через 1,5 - 2,0 часа от начала рабочей смены и через 1,5 - 2,0 часа после обеденного перерыва продолжительностью 20 минут каждый или продолжительностью 15 минут через каждый час работы.

5.3. Экологическая безопасность

Целью данного подраздела является выявление потенциальных опасностей объекта и процесса исследования на окружающую среду, а также разработка мер, обеспечивающая безопасность исследовательской деятельность для окружающей среды.

5.3.1. Анализ влияния объекта исследования на окружающую среду

Объект исследования (методология подготовки исходных данных) не оказывает влияния на окружающую среду, поскольку используется только совместно с ПЭВМ. Сами ПЭВМ могут являться источниками различных загрязнений окружающей среды.

5.3.2. Анализ влияния процесса исследования на окружающую среду

Процесс исследования включает в себя работу на ПЭВМ в учебной аудитории (КЦ ТПУ), в том числе в условиях искусственного освещения, обеспечиваемого люминесцентными лампами.

Отработанная офисная техника относится к опасным отходам. При производстве компьютеров и других агрегатов применяются вещества, опасные для жизнедеятельности, например, свинец, мышьяк и др. Обычное выбрасывание техники, особенно регулярное, может нанести непоправимый вред экологии и здоровью. Согласно Административному Кодексу РФ, ст. 8.2 [18], запрещается выбрасывать технику наряду с обыкновенным мусором, причем запрет распространяется не только на физических лиц, но и на организации.

5.3.3. Обоснование мероприятий по защите окружающей среды

Согласно ГОСТ Р 56397-2015 [19], в результате технической экспертизы может быть принято следующее решение: оборудование не ремонтопригодно, признается неработоспособным и рекомендуется к списанию (замене); в случае деградационного отказа оборудования и нецелесообразности его ремонта и модернизации даются рекомендации о необходимости его списания и утилизации. Самостоятельная утилизация оргтехники запрещена, утилизация производится только в промышленных условиях. Утилизировать компьютерную технику имеют права специализированные предприятия при наличии соответствующей лицензии.

На рабочем месте программиста используются 16 люминесцентных ламп ЛБ40, Согласно ГОСТ 12.3.031-83 [20] «Работы со ртутью. Требования безопасности» п.2.1. все ртутьсодержащие отходы и вышедшие из строя приборы, содержащих ртуть, подлежат сбору и возврату для последующей регенерации ртути в специализированных организациях. В п.2.2. К работе по замене и сбору отработанных ртутьсодержащих ламп допускаются только электромонтеры. Главным условием при замене и сборе отработанных ртутьсодержащих ламп является сохранение герметичности. В п.2.13. Факт сдачи ртутьсодержащих отходов подтверждается возвращением паспорта на вывоз отходов с отметкой о приеме представителя специализированного предприятия.

5.4. Безопасность в чрезвычайных ситуациях

5.4.1. Анализ вероятных ЧС, которые может инициировать объект исследований

В ходе проведения анализа не было выявлено ЧС, которые может инициировать объект исследования напрямую.

5.4.2. Анализ вероятных ЧС, которые могут возникнуть на рабочем месте при проведении исследований

К наиболее вероятным ЧС на рабочем месте можно отнести следующие: пожар (взрыв) в здании, авария на коммунальных системах жизнеобеспечения, землетрясение.

Наиболее вероятным ЧС является пожар. Источниками возгорания может стать электропроводка, внутренние работающие устройства ПК, взрывоопасные предметы в помещении исследователя согласно ГОСТ 12.1.044-2018 «Система стандартов безопасности труда. Пожаровзрывоопасность веществ и материалов. Номенклатура показателей и методы их определения» [21].

Поражающими факторами пожаров в помещении являются токсическое воздействие горючих материалов (в т.ч. отравление угарным газом), экстремальный нагрев среды, а также обломки и осколки при нарушении целостности конструкций здания [22].

5.4.3. Обоснование мероприятий по предотвращению ЧС и разработка порядка действий в случае возникновения ЧС

Согласно ГОСТ Р 22.3.03-94 [23], обеспечение безопасности людей в ЧС, обусловленных природными стихийными бедствиями, техногенными авариями и катастрофами, а также применением современного оружия (военные ЧС) является общегосударственной задачей, обязательной для решения всеми территориальными, ведомственными и функциональными органами управления и регулирования, службами и формированиями, а также подсистемами, входящими в Российскую систему предупреждения и действий в чрезвычайных ситуациях (РСЧС).

Мероприятия по защите людей от источников ЧС должны планироваться в объемах, гарантирующих непревышение нормативного воздействия на них возможных поражающих факторов для расчетной ЧС.

Для защиты жизни и здоровья населения в ЧС следует применять следующие основные мероприятия гражданской обороны, являющиеся составной частью мероприятий РСЧС:

1) укрытие людей в приспособленных под нужды защиты населения помещениях производственных, общественных и жилых зданий, а также в специальных защитных сооружениях;

2) эвакуацию населения из зон ЧС;

3) использование средств индивидуальной защиты органов дыхания и кожных покровов;

4) проведение мероприятий медицинской защиты;

5) проведение аварийно-спасательных и других неотложных работ в зонах ЧС.

Мерами по предупреждению ЧС являются:

1) Соблюдение техники безопасности при работе с ПЭВМ. Использование только исправного оборудования.

2) Своевременное проведение ТО и ППР электроустановок согласно утвержденного графика и технических средств противопожарной защиты и пожаротушения.

3) Установка противопожарной сигнализации.

4) Своевременное проведение инструктажа рабочего персонала.

В случае угрозы возникновения ЧС (пожара) необходимо вызвать противопожарную службу, отключить электроэнергию и, следуя плану эвакуации, эвакуировать находящихся в помещении людей и покинуть помещение. В случае, если очаг возгорания является небольшим, и нет угрозы поражения электрическим током, можно использовать углекислотные огнетушители ОУ-5 высокого давления с зарядом жидкой двуокиси углерода, согласно ГОСТ 8050-85 [24]. Расположение огнетушителей отмечено на плане эвакуации людей при пожаре и других ЧС из помещений учебного корпуса №22.

5.5. Вывод по главе

В данной главе были проанализированы опасные и вредные факторы труда разработчика, а также предложены меры защиты от них, оценены условия труда рабочей зоны; рассмотрены требования по технике безопасности, электробезопасности, пожарной безопасности, экологической безопасности. В результате анализа было установлено, что аудитория Кибернетического центра ТПУ удовлетворяет всем требованиям, предъявляемым к нему нормативными документами в области охраны труда и окружающей природной среды.

Заключение

В результате выполнения магистерской диссертации были выполнены следующие задачи:

1. Устранение ошибок и восстановление пропущенных значений с помощью алгоритма поиска ближайшего соседа (k-nearest algorithm imputation).
2. Построение корреляционной матрицы для исключения линейной зависимости между атрибутами
3. Определение целевой функции для построения модели машинного обучения
4. Проведение регрессионного анализа
5. Выделение признаков по важности

В результате проведения разведочного и регрессионного анализа можно сделать вывод о том, что на предсказание о количестве добываемой нефти в рассматриваемом исследовании значительно влияют лишь 10 параметров (в начале исследования их было порядка 100 единиц).

В ходе работы был получен опыт работы с такими языками программирования как Python и R, средами разработки Jupyter Notebook и R Studio, а также были использованы следующие инструменты и библиотеки: pandas, numpy, seaborn и несколько библиотек из среды разработки R Studio.

Список публикаций и научных достижений

Участие в конференциях:

1. XVl Международная научно-практическая конференция студентов, аспирантов и молодых ученых «Молодёжь и современные и информационные технологии» с докладами:

- Анализ социальных данных с помощью технологий Big Data
- Проектирование виртуального полигона для беспилотного автомобиля
- Разработка виртуального полигона для беспилотного автомобиля

2. VI Международная научная конференция «Информационные технологии в науке, управлении, социальной сфере и медицине» с докладом:

- Подготовка исходных данных для углубленного анализа нефтегазового месторождения

Премии, звания, стипендии:

1. Повышенная государственная стипендия по научно-исследовательской деятельности, осень 2019/2020 учебного года, весна 2019/2020 учебного года.
2. Стипендия Президента РФ (или Правительства РФ) студентам ТПУ, обучающимся по специальностям или направлениям подготовки, соответствующим приоритетным направлениям модернизации и технологического развития российской экономики, осень, осень 2019/2020 учебного года.

Публикации:

1. Журбич Н. И. Подготовка исходных данных для углубленного анализа нефтегазового месторождения // Информационные технологии в науке, управлении, социальной сфере и медицине: сборник научных трудов VI Международной научной конференции, Томск, 14-19 Октября 2019. - Томск: ТПУ, 2019 - С. 13-19
2. Журбич Н. И. Проектирование виртуального полигона в Unity3D // Информационные технологии в науке, управлении, социальной сфере и медицине: сборник научных трудов V Международной конференции: в 2 т., Томск, 17-21 Декабря 2018. - Томск: ТПУ, 2018 - Т. 1 - С. 249-251.

3. Журбич Н. И. Проектирование виртуального полигона для беспилотного автомобиля // Молодежь и современные информационные технологии: сборник трудов XVI Международной научно- практической конференции студентов, аспирантов и молодых ученых, Томск, 3-7 Декабря 2018. - Томск: ТПУ, 2019 - С. 427-428.

4. Журбич Н. И., Фофанов О. Б. Разработка виртуального полигона для беспилотного автомобиля // Молодежь и современные информационные технологии: сборник трудов XVI Международной научно- практической конференции студентов, аспирантов и молодых ученых, Томск, 3-7 Декабря 2018. - Томск: ТПУ, 2019 - С. 429-430.

5. Журбич Н. И., Зяблецев П. А. Анализ данных с помощью технологий Big Data // Информационные технологии в науке, управлении, социальной сфере и медицине: сборник научных трудов V Международной конференции: в 2 т., Томск, 17-21 Декабря 2018. - Томск: ТПУ, 2018 - Т. 1 - С. 255-257.

6. Журбич Н. И., Зяблецев П. А. Анализ социальных данных с помощью технологий Big Data // Молодежь и современные информационные технологии: сборник трудов XVI Международной научно- практической конференции студентов, аспирантов и молодых ученых, Томск, 3-7 Декабря 2018. - Томск: ТПУ, 2019 - С. 148-149.

7. Журбич Н. И., Зяблецев П. А. Разработка виртуального полигона в Unity 3D // Информационные технологии в науке, управлении, социальной сфере и медицине: сборник научных трудов V Международной конференции: в 2 т., Томск, 17-21 Декабря 2018. - Томск: ТПУ, 2018 - Т. 1 - С. 252-255

8. Зяблецев П. А., Журбич Н. И. Выявление факторов риска острого инфаркта миокарда с помощью OLAP технологии // Информационные технологии в науке, управлении, социальной сфере и медицине: сборник научных трудов V Международной конференции: в 2 т., Томск, 17-21 Декабря 2018. - Томск: ТПУ, 2018 - Т. 1 - С. 267-270.

Список используемых источников

1. Лидерами нефтегаза станут компании, использующие Big Data. [Электронный ресурс]. CNEWS. URL: http://www.cnews.ru/news/top/liderami_neftegaza_stanut_kompanii (дата обращения: 03.09.2019).
2. Перспективные технологии Big Data в нефтяном инжиниринге: опыт компании «Газпром Нефть». [Электронный ресурс]. НТЦ Газпром. URL: https://ntc.gazprom-neft.ru/research-and-development/papers/13596/ (дата обращения: 03.05.2020).
3. Как Big Data и Machine Learning в нефтегазовой отрасли помогают экономить миллиарды [Электронный ресурс]. Школа больших данных. URL: https://www.bigdataschool.ru/bigdata/machine-learning (дата обращения: 08.05.2020).
4. Губин Е.И. Методика подготовки больших данных для прогнозного анализа //Наука и бизнес: Пути развития, № 3(105) 2020, с. 27-31.
5. ЭЦН. [Электронный ресурс]. Википедия. URL: https://ru.wikipedia.org/wiki/%D0%AD%D0%A6%D0%9D (дата обращения: 21.08.2019).
6. Журбич Н. И. Подготовка исходных данных для углубленного анализа нефтегазового месторождения // Информационные технологии в науке, управлении, социальной сфере и медицине: сборник научных трудов VI Международной научной конференции, Томск, 14-19 Октября 2019. - Томск: ТПУ, 2019 - С. 13-19.
7. Геолого-технические мероприятия (ГТМ). [Электронный ресурс]. CNEWS. URL: https://www.petroleumengineers.ru/forum/39 (дата обращения: 27.08.2019).
8. The use of KNN for missing values. [Электронный ресурс]. Towards Data Science. URL: https://towardsdatascience.com/the-use-of-knn-for-missing-values-cf33d935c637 (дата обращения: 03.09.2019).

9. Трудовой кодекс Российской Федерации от 30.12.2001 N 197-ФЗ (ред. от 01.04.2019).

10. Федеральный Закон от 27.07.2006 N 152-ФЗ (ред. от 25.07.2011) «О Персональных Данных».

11. ГОСТ 12.2.032-78 ССБТ. Рабочее место при выполнении работ сидя. Общие эргономические требования.

12. СанПиН 2.2.2/2.4.1340-03 Гигиенические требования к персональным электронно-вычислительным машинам и организации работы (с изменениями на 21 июня 2016 года)

13. ГОСТ 12.0.003-2015 Система стандартов безопасности труда (ССБТ). Опасные и вредные производственные факторы. Классификация.

14. СанПиН 2.2.4.548-96 Гигиенические требования к микроклимату производственных помещений.

15. ГОСТ 12.1.003-2014 Система стандартов безопасности труда (ССБТ). Шум. Общие требования безопасности.

16. Безопасность жизнедеятельности: практикум / Ю.В. Бородин, М.В. Василевский, А.Г. Дашковский, О.Б. Назаренко, Ю.Ф. Свиридов, Н.А. Чулков, Ю.М. Федорчук. — Томск: Изд-во Томского политехнического университета, 2009. — 101 с.

17. ГОСТ 12.1.019-2017 Система стандартов безопасности труда (ССБТ). Электробезопасность. Общие требования и номенклатура видов защиты.

18. ТОИ Р-45-084-01 Типовая инструкция по охране труда при работе на персональном компьютере.

19. ГОСТ 12.1.029-80 Система стандартов безопасности труда (ССБТ). Средства и методы защиты от шума. Классификация.

20. КоАП РФ Статья 8.2 Несоблюдение экологических и санитарно-эпидемиологических требований при обращении с отходами производства и потребления, веществами, разрушающими озоновый слой, или иными опасными веществами.

21. ГОСТ Р 56397-2015 Техническая экспертиза работоспособности радиоэлектронной аппаратуры, оборудования информационных технологий, электрических машин и приборов. Общие требования.

22. ГОСТ 4658-73 Ртуть. Технические условия (с Изменениями N 1-6).

23. ГОСТ 12.1.044-2018 Система стандартов безопасности труда. Пожаровзрывоопасность веществ и материалов. Номенклатура показателей и методы их определения.

24. ГОСТ Р 22.0.07-95 Безопасность в чрезвычайных ситуациях. Источники техногенных чрезвычайных ситуаций. Классификация и номенклатура поражающих факторов и их параметров.

25. ГОСТ Р 22.3.03-94. Безопасность в чрезвычайных ситуациях. Защита населения. Основные положения.

26. ГОСТ 8050-85 Двуокись углерода газообразная и жидкая. Технические условия (с Изменениями N 1, 2, с Поправкой).

Приложение 1 (рус.яз)
(справочное)

Using big data technologies in the oil and gas sector

Студент:

Группа	ФИО	Подпись	Дата
8ПМ8И	Журбич Никита Игоревич		05.06.2020

Руководитель ВКР:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОИТ ИШИТР	Губин Евгений Иванович	к.ф. – м.н.		05.06.2020

Консультант – лингвист отделения иностранных языков ШБИП:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Старший преподаватель ОИЯ	Пичугова Инна Леонидовна	-		05.06.2020

Introduction

Oil and gas companies in their activities receive petabytes of data every day. The use of big data opens up the possibility of analyzing and predicting the development of trends in the field of geology, engineering, production and the best way to use equipment to achieve the most optimal results at all stages of their activities.

Oil and gas companies will not be able to take advantage of Big Data's competitive advantage unless they manage their data more efficiently. This conclusion was made in new report by the oil and gas consulting company Molten. According to the experts, many oil and gas companies "irresponsibly" manage their data, despite the fact that they spend billions of dollars a year on their collection. According to Molten estimates, large oil and gas companies spend from $ 1 to $ 3 billion per year on data collection, but the costs of maintaining and processing the accumulated information often make up less than 1% of this amount. At the same time, companies are required to make operational decisions and maintain a high level of productivity. As a result, management must rely on large amounts of data to make critical decisions. The scope of application of Big Data technology in the oil and gas industry is very extensive, and includes the whole spectrum, from exploration and development to hydrocarbon processing.

The object of research is the process of developing a methodology for processing source data from an oil and gas field to build a predictive model.

The purpose of this master's thesis is to develop a methodology for processing source data from an oil field to build a predictive model.

1. Domain review

1.1. Fields of application of Big Data technologies in oil engineering

Currently, Big Data technologies are one of the key drivers of information technology development. However, there are few successful cases in world oil engineering. This is due to the historically inertia characteristic of many fundamental industries. Over a period of about 10 years of high oil prices, information technology was not seen as a significant growth driver in the oil industry. Modern realities dictate necessity to optimize processes and increase operational efficiency in the oil and gas sector.

A large number of successful projects have been implemented in the field of data processing automation, for example, in projects for the creation of digital deposits, and in predictive analytics to assess the reliability and predict complications in the operation of equipment in various technological processes, mainly in drilling. According to estimates, the introduction of predictive analytics systems based on big data analysis in drilling can reduce well construction time by 30%, and the total cost of a well, including development, by 15%. In addition, with the help of Big Data tools, a wide range of tasks in logistics is successfully solved: from optimizing transport routes and equipment supply schemes to increasing the efficiency of gas stations [1].

In Figure 1, the priorities for introducing Big Data technologies in the oil and gas sector were highlighted.

The distributed structure and high capital intensity of the industry lead to the daily appearance of a multitude diverse data that must be collected and analyzed in order to reduce current expenses and increase future profits. This is an incentive for introducing Big Data and Machine Learning technologies in the oil and gas industry into production processes, in particular, predictive or predictive analytics systems.

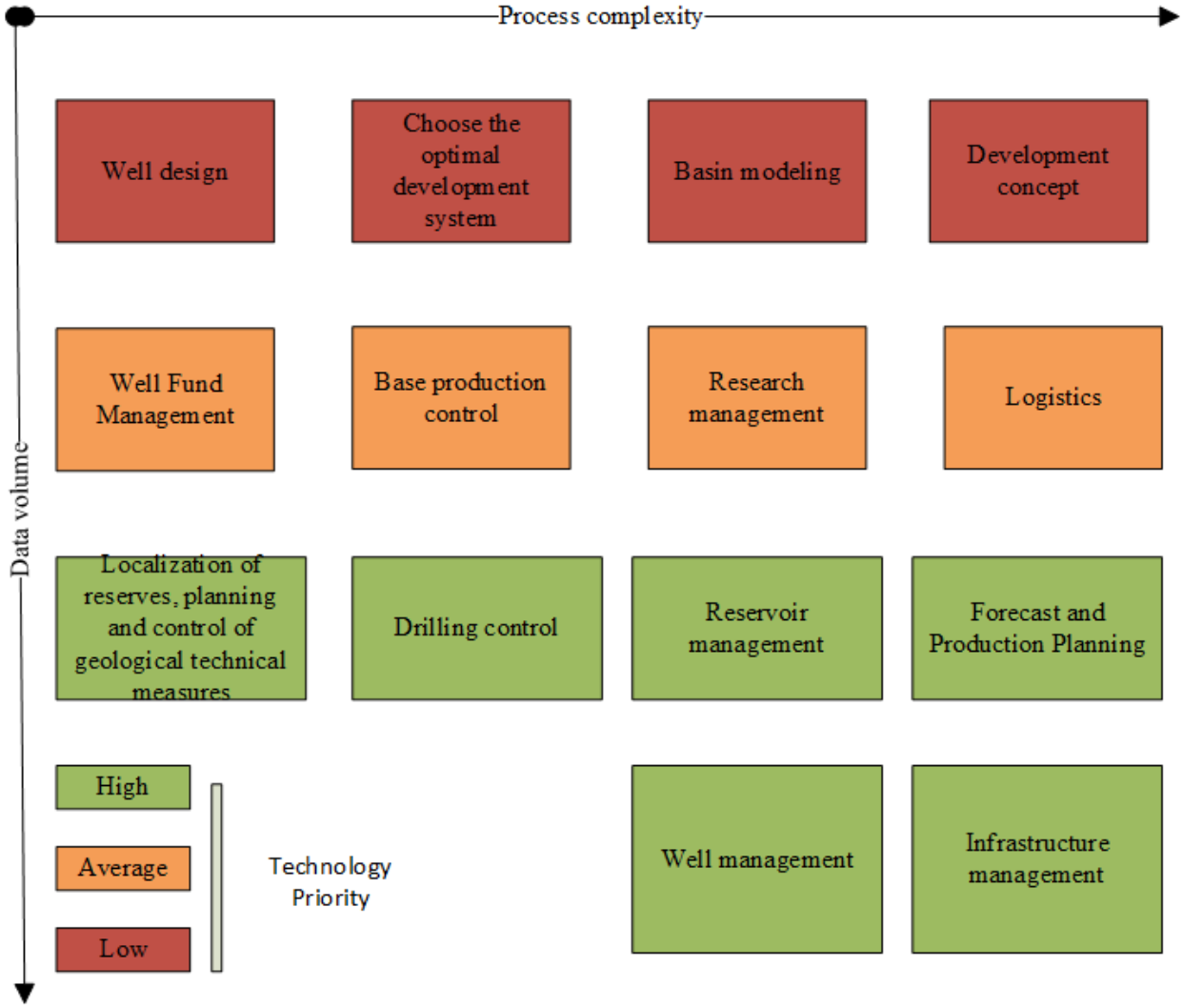


Figure 1. Priorities for introducing Big Data technologies in the oil and gas sector

Based on the presented priorities for introducing Big Data technologies in the oil and gas sector, the most priority tasks such as well management, infrastructure management, forecast and production planning and reservoir management can be noted. In this paper, one of the priority tasks will be the development of a machine learning model for predicting certain parameters.

1.2. Problem and purpose

The oil and gas industry in Russia is the main source of foreign exchange and tax revenues of the country. It accounts for about 12% of all industrial production and more than 40% of budget revenues. However, despite the almost century-old domestic history of this industry, its current state is accompanied by many problems that the latest information technologies should solve.

To study and search for the causes of the problem under consideration, the Fish Skeleton diagram is used. This diagram allows you to identify all potential causes of the problem in a simple and affordable way.

The following problem was chosen as the main problem: "Lack of analogues of such a system on the market." Figure 2 shows the Fish Skeleton diagram for analyzing the subject area and finding the cause.

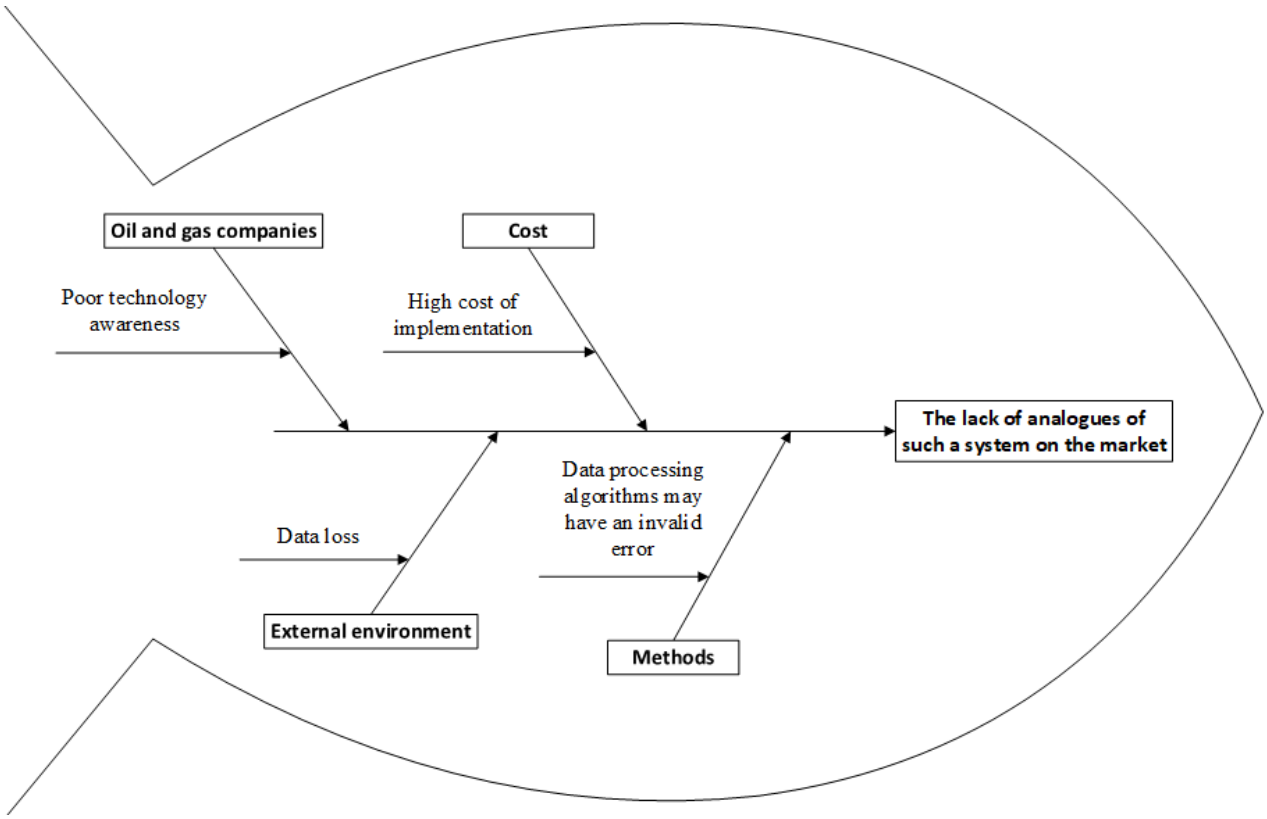


Figure 2. The diagram "Fish skeleton" for the subject area

The problem is the lack of analogues of the methodology in question in the market. Most oil and gas companies are just now starting to use big data analysis tools, despite the success of these tools in other areas of production.

The goal is to develop a methodology that will allow for the correct exploratory data analysis.

1.3. Experience in using Big Data technologies in the oil and gas industry

Foreign experience in using Machine Learning technologies in the oil and gas industry:

1. Reduction in the construction time of wells by 30%, reduction in the total cost of the well − by 15%

The British company British Petroleum has managed to reduce equipment operating costs by more than $ 2 million thanks to well prediction systems for electric centrifugal pumps on offshore platforms. This has contributed to increasing the period of operation of the wells and reducing downtime in anticipation of equipment repair.

2. Optimization of diagnostic checks of oil and gas equipment

The American company GE, General Electric Oil & Gas (specializing in the production of equipment for the oil and gas field) has implemented machine learning algorithms in one of divisions. Automated analysis of the collected data allowed the company's engineers to optimize the schedule of diagnostic checks, improve the efficiency of equipment use and reduce downtime by proactively identifying possible malfunctions. Because of this, there is an increase in annual energy production and a decrease in losses from the inefficient use of technology [2].

3. Reducing the cost of oil production due to the concept of "digital field"

A digital field refers to assets that are equipped with a set of monitoring and remote control systems, as well as specialized software for production processes. The approach is aimed at increasing oil and gas production, as well as reducing downtime and labor costs by optimizing work and reducing shortfalls. Digital

fields provide an optimal technological mode of oil production, which allows reducing the cost of production by 7-10%, and the cost of operating a field facility by 20%.

Shell, Chevron, BP and Schlumberger actively use this approach not only for continuous monitoring of the state of their technological systems and processes, but also for predictive modeling of expected and emergency situations in the short and long term.

Domestic experience in using Machine Learning technologies in the oil and gas industry:

1. Identify the causes of failures in the automatic restart of the pumps after an emergency power outage

Gazpromneft analyzed more than 200 million various records received during the year from controllers of control systems at 1,649 wells, records of voltage restarts from emergency logs and factors of the dependence of pump operation on well conditions, operation features, power supply schemes, etc.

Analytical tools made it possible to formulate and test a set of hypotheses about the causes of failures and obtain information on previously unknown relationships in the operation of pumping equipment, for example, the appearance of a turbine rotation effect, which leads to a reverse oil drain when the pump power is turned off [3].

2. Modeling production facilities and launching digital fields

At present, Gazpromneft is exploring algorithms for automated selection of the optimal system for the development of newly commissioned fields and optimization of well operation modes at long-developed facilities to maximize production. This will make it possible at the planning stages to select the most effective technology for the development of deposits, and during their further operation − to increase the efficiency of developing residual recoverable reserves. The potential economic effect is estimated at 1 million tons of additional production. A project has also been launched for the intelligent search for analogous objects according to a given set of criteria using machine learning, the

implementation of which by 2025 will reduce the company's costs by 4 billion rubles.

A project is aimed to search for missed intervals according to the data of geophysical surveys of wells, which can bring about 500 thousand tons of additional production from current production assets. An expert analysis of the missed intervals showed that the digital model of the field allows us to allocate 14% more additional effective thicknesses than the current results of interpretation of the data of the used geographic information systems show.

3. Gas Theft Detection

In 2018, Gazprom developed a new analytical algorithm for installation in all of its subsidiary gas supply and gas distribution companies. For example, now the theft problem is most acute in the North Caucasus Federal District, where gas losses reach 3.5 billion cubic meters per year, i.e. 16 billion rubles. It is planned to reduce the imbalance of recorded and unaccounted fuel due to the continuous collection of consumption data, their systematization and monitoring of consumption by balance zones. Projected traceability − at least 90% of total consumption.

All of these technologies are aimed at improving the efficiency of existing processes and creating a technological backlog in Gazprom Neft. The introduction of these technologies will allow the business to receive a number of unique advantages.

1. Improving the quality and timeliness of production decisions based on geological and hydrodynamic models (GDM) by improving the quality of digital models, reducing the duration of the GDM cycle and minimizing the influence of the "human factor" in the interpretation of studies.

2. Improving the validity and quality of investment decision-making in conditions of ultra-high uncertainty in the source data, and often their lack.

1.4. Advantages and disadvantages of developing a methodology

Big data processing is changing almost all areas of our lives, and business in the first place. However, for some reason, not all enterprises, even large ones, have switched to using this technology.

Table 1 summarizes the main advantages and disadvantages of developing a methodology for processing data obtained from an oil and gas field.

Table 1. Main advantages and disadvantages of data processing methodology

Advantages	Disadvantages
Increased productivity − modern tools are allowing employees to analyze more data, which increases their personal productivity	**Data quality** − working with big data was the need to address data quality issues.
Better decision-making − analytics can give business decision-makers the data-driven insights they need to help their companies compete and grow.	**Compliance** − another issue for big analytics efforts is complying with government regulations.
Fraud detection − banks and credit card companies may spot stolen credit cards or fraudulent purchase	**Cybersecurity risks** − storing sensitive data can make companies a more attractive target for cyberattackers.
Reduce costs − big data analytics help companies decrease their expenses.	**Hardware needs** − expensive hardware

Improved customer service − social media give today's enterprises a wealth of information about their customers, and it is only natural that they would use this data to better serve those customers.	**Need for talent** − data scientists and big data experts are among the most highly coveted and highly paid workers in the IT field
Increased revenue − when organizations use big data to improve their decision-making and improve their customer service, increased revenue is often the natural result.	**Need for cultural change** − many of the organizations that are utilizing big data analytics do not just want to get a little bit better at reporting, they want to use analytics to create a data-driven culture throughout the company.
Increased agility − many organizations are using their big data to better align their IT and business efforts, and they are using their analytics to support faster and more frequent changes to their business strategies and tactics.	**Difficulty integrating legacy systems** − integrating all disparate data sources and moving data where it needs to be also adds to the time and expense of working with big data.

Despite the increased requirements for software and hardware and the high cost of storing and processing data, the biggest drawback is the problem of choosing the data to be processed, i.e. it is necessary to determine what data needs to be extracted, stored and analyzed, and which data needs to be deleted.

1.5. Conclusion

The problem of processing and storing big data in oil and gas companies remains relevant to this day for several reasons. First of all, it is the impossibility of using traditional approaches. Despite the variety of technologies and methods for solving this problem, there is no universal way to process a large amount of data obtained from oil and gas fields. This conclusion emphasizes the need to create a methodology for processing this data for further research and experimentation in the oil and gas sector.

Thus, no modern company can do without the processing of digital information, and the growth rate of the volume of information is increasing every day. The quality of processing and completeness of information is the key to making the right management decisions. Those companies that correctly understand the problem and adequately deal with it, including developing and implementing Big Data technologies, will remain afloat a large flow of information.

Conclusion

In the process of completing the master's thesis, the following tasks were performed:

- Elimination of errors and restoration of missing values using the k-nearest algorithm imputation algorithm.
- Construction of a correlation matrix to exclude a linear relationship between attributes
- The definition of the objective function to build a machine learning model
- Conducting regression analysis
- Construction and training of a regressor obtained by random forest method;
- Assessment of the accuracy of each model by certain metrics;
- Identify the most important features for the selected objective function.

As the result of exploration and regression analysis, we can conclude that only 10 parameters significantly affect the prediction of the amount of oil produced in the study in question (there were about 100 units at the beginning of the study).

In the course of work, I gained experience working with programming languages such as Python and R, the Jupyter Notebook and R Studio development environments, and used the following tools and libraries: pandas, numpy, seaborn and several libraries from the R Studio development environment.

References

1. How big data is changing the oil & gas industry? URL: http://analytics-magazine.org/how-big-data-is-changing-the-oil-a-gas-industry/
2. Big Data analytics in oil and gas URL: https://www.bain.com/contentassets/38df79e1ec42486497d197c48e09118b/bain_brief_big_data_in_oil_and_gas.pdf
3. How to use Big Data technologies to optimize operations in Upstream Petroleum Industry. URL: https://hal.archives-ouvertes.fr/hal-00944668/file/Paper_IJI_Baaziz_Quoniam.pdf

Приложение 2

Скрипт на Python

```
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import seaborn as sns
import os
neft1 = pd.read_csv('neft1.csv',sep = ";")
neft1.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3367 entries, 0 to 3366
Columns: 123 entries, Скважина to ГП - Общий прирост Qн
dtypes: float64(66), object(57)
memory usage: 2.4+ MB

neft1.head(5)
## to check which columns have null values.
neft1.isna().any()
Скважина                        False
Дата                           False
ГТМ                             True
Метод                           True
Характер работы                 True
Состояние                       True
Время работы, ч                 True
Время накопления                True
Нефть, т                        True
Попутный газ, м3                True
Закачка, м3                     True
Природный газ, м3               True
Газ из газовой шапки, м3        True
Конденсат, т                    True
Простой, ч                      True
Причина простоя                 True
Приемистость, м3/сут            True
Обводненность (вес), %          True
```

Агент закачки True

Нефть, м3 True

Жидкость, м3 True

Дебит конденсата True

Добыча растворенного газа, м3 True

Дебит попутного газа, м3/сут True

Пласт МЭР True

Куст True

Тип скважины True

Диаметр экспл.колонны True

Диаметр НКТ True

Диаметр штуцера True

...

Глубина текущего забоя True

Тип дополнительного оборудования True

Диаметр дополнительного оборудования True

Глубина спуска доп. оборудования True

Тип газосепаратора True

Марка ПЭД True

Мощность ПЭД True

I Х/Х True

Ток номинальный True

Ток рабочий True

Число качаний ШГН True

Длина хода плунжера ШГН True

Диаметр плунжера True

Станок-качалка True

Коэффициент подачи насоса True

Тип ГЗУ True

ДНС True

КНС True

КН закрепленный True

Пластовое давление начальное True

Характеристический дебит жидкости True

Время в работе True

Время в накоплении True
ГП - Забойное давление True
ГП(ИДН) Дебит жидкости True
ГП(ИДН) Дебит жидкости скорр-ый True
ГП(ИДН) Прирост дефита нефти True
ГП(ГРП) Дебит жидкости True
ГП(ГРП) Дебит жидкости скорр-ый True
ГП - Общий прирост Qн True
Length: 123, dtype: bool

```
import seaborn as sns
sns.countplot(neft1['Состояние'])
sns.countplot(neft1['ГТМ'])
neft1.describe()
neft1.isnull().sum()
```

Скважина 0
Дата 0
ГТМ 461
Метод 544
Характер работы 469
Состояние 480
Время работы, ч 461
Время накопления 461
Нефть, т 461
Попутный газ, м3 461
Закачка, м3 461
Природный газ, м3 461
Газ из газовой шапки, м3 461
Конденсат, т 461
Простой, ч 461
Причина простоя 2908
Приемистость, м3/сут 461
Обводненность (вес), % 463
Агент закачки 3277
Нефть, м3 461

Жидкость, м3 471

Дебит конденсата 471

Добыча растворенного газа, м3 463

Дебит попутного газа, м3/сут 463

Пласт МЭР 2128

Куст 2143

Тип скважины 2352

Диаметр экспл.колонны 2128

Диаметр НКТ 2128

Диаметр штуцера 2138

...

Глубина текущего забоя 2148

Тип дополнительного оборудования 2901

Диаметр дополнительного оборудования 2148

Глубина спуска доп. оборудования 2148

Тип газосепаратора 3367

Марка ПЭД 3025

Мощность ПЭД 2148

I X/X 2148

Ток номинальный 2148

Ток рабочий 2148

Число качаний ШГН 2148

Длина хода плунжера ШГН 2148

Диаметр плунжера 2148

Станок-качалка 3367

Коэффициент подачи насоса 2148

Тип ГЗУ 2148

ДНС 2148

КНС 2827

КН закрепленный 2148

Пластовое давление начальное 2148

Характеристический дебит жидкости 2148

Время в работе 2148

Время в накоплении 2148

ГП - Забойное давление 2148

```
ГП(ИДН) Дебит жидкости                2148
ГП(ИДН) Дебит жидкости скорр-ый       2148
ГП(ИДН) Прирост дефита нефти          2148
ГП(ГРП) Дебит жидкости                2148
ГП(ГРП) Дебит жидкости скорр-ый       2148
ГП - Общий прирост Qн                 2148
Length: 123, dtype: int64
```

```
import matplotlib.pyplot as plt
import seaborn as sns
# correlation visualization triangle
corr = df.corr()
# Plot figsize
fig, ax = plt.subplots(figsize=(20, 20))
dropSelf = np.zeros_like(corr)
dropSelf[np.triu_indices_from(dropSelf)] = True
# Generate Color Map
colormap = sns.light_palette((210, 90, 60), input="husl")
# Generate Heat Map, allow annotations and place floats in map
sns.heatmap(corr, cmap=colormap, annot=True, fmt=".2f", mask=dropSelf)
# Apply xticks
plt.xticks(range(len(corr.columns)), corr.columns);
# Apply yticks
plt.yticks(range(len(corr.columns)), corr.columns)
plt.show()
```

Приложение 3

Скрипт на R

```
> neft <- read.csv(file="C:/neft2.csv", header=TRUE, sep=";",na.strings=c("",
" ","NA"))
> head(neft)
                          Скважина       Дата ГТМ   Метод Характер.работы Сос
тояние Время.работы..ч
1 002ff5b8a6dc271f58581e1b4fa2c5fc 01.12.2016   1     ФОН             НЕФ
ОСВ ТГ               0
2 008d0347e572a5d938a9c40c29e539fc 01.10.2013  NA    <NA>            <NA>
  <NA>               NA
3 00b40cb7bb8c9fd1ac26b4cc86f2b291 01.02.2018  NA    <NA>            <NA>
  <NA>               NA
4 01ba18d8b6d29875a18d4bca4eb201d7 01.05.2014   0 ЭЦН/ФОН             НЕФ
  РАБ.             120
5 024ec6f6e3f9c5150ecf525bf8b7a6a3 01.06.2017   1     ФОН             НЕФ
ОСВ ТГ               0
6 0254a227c6c2c31a419126700cfcddc2 01.05.2017   1 ЭЦН/ФОН             НЕФ
  ОСТ.
> imputed_dat <- kNN(neft)
> head(imputed_dat)
                          Скважина       Дата ГТМ   Метод Характер.работы Сос
тояние Время.работы..ч
1 002ff5b8a6dc271f58581e1b4fa2c5fc 01.12.2016   1     ФОН             НЕФ
ОСВ ТГ               0
2 008d0347e572a5d938a9c40c29e539fc 01.10.2013   0     ЭЦН             НЕФ
  РАБ.             384
3 00b40cb7bb8c9fd1ac26b4cc86f2b291 01.02.2018   0     ЭЦН             НЕФ
  РАБ.             384
4 01ba18d8b6d29875a18d4bca4eb201d7 01.05.2014   0 ЭЦН/ФОН             НЕФ
  РАБ.             120
5 024ec6f6e3f9c5150ecf525bf8b7a6a3 01.06.2017   1     ФОН             НЕФ
ОСВ ТГ               0
6 0254a227c6c2c31a419126700cfcddc2 01.05.2017   1 ЭЦН/ФОН             НЕФ
  ОСТ.
> aggr(imputed_dat, prop = F, numbers = T,sortVars=TRUE)

 Variables sorted by number of missings:
                        Variable Count
                        Скважина     0
                            Дата     0
                             ГТМ     0
                           Метод     0
                 Характер.работы     0
                       Состояние     0
                 Время.работы..ч     0
                Попутный.газ..м3     0
                      Простой..ч     0
```

```
                    Причина.простоя     0
             Обводненность..вес....     0
      Добыча.растворенного.газа..м3     0
       Дебит.попутного.газа..м3.сут     0
                       Скважина_imp     0
                           Дата_imp     0
                            ГТМ_imp     0
                          Метод_imp     0
                 Характер.работы_imp     0
                       Состояние_imp     0
                 Время.работы..ч_imp     0
               Попутный.газ..м3_imp     0
                     Простой..ч_imp     0
                 Причина.простоя_imp     0
        Обводненность..вес...._imp     0
 Добыча.растворенного.газа..м3_imp     0
  Дебит.попутного.газа..м3.сут_imp     0
> aggr(neft, prop = F, numbers = T,sortVars=TRUE)

 Variables sorted by number of missings:
                       Variable Count
                Причина.простоя   215
                            ГТМ    45
                          Метод    45
                 Характер.работы    45
                       Состояние    45
                 Время.работы..ч    45
               Попутный.газ..м3    45
                     Простой..ч    45
        Обводненность..вес....     45
 Добыча.растворенного.газа..м3     45
  Дебит.попутного.газа..м3.сут     45
                       Скважина     0
                           Дата     0
```

Приложение 4

Скрипт на python

```
#Regression for MPG (1 parameter only)
import numpy as np
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
train_df, test_df = train_test_split(df, test_size = 0.3, random_state = 0)
X = train_df[['Нефть..м3']].values
y = train_df['Дебит.жидкости'].values
model = LinearRegression()
model.fit(X, y)
y_pred = model.predict(X)
plt.scatter(X, y)
plt.plot(X, model.predict(X), color='red', linewidth=2)
plt.xlabel('Нефть,м3')
plt.ylabel('Дебит жидкости')
from sklearn.metrics import mean_absolute_error, mean_squared_error, median_absolute_error,
r2_score
print('MSE train: {:.3f}, test: {:.3f}'.format(
      mean_squared_error(y_train, y_pred_train),
      mean_squared_error(y_test, y_pred_test)))
print('R^2 train: {:.3f}, test: {:.3f}'.format(
      r2_score(y_train, y_pred_train),
      r2_score(y_test, y_pred_test)))
#Regression for multiple parameters
from sklearn.model_selection import train_test_split
Y = df['Нефть..м3'].values
X = df[['Дебит.жидкости','Попутный.газ..м3'
,'Жидкость..м3','Добыча.растворенного.газа..м3', 'Дебит.попутного.газа..м3.сут', ]]
x_train, x_test, y_train, y_test = train_test_split(X, Y, test_size=0.3, random_state=0)
from sklearn.neural_network import MLPRegressor
regr = MLPRegressor(activation='relu', solver='adam', random_state=0)
regr.fit(x_train, y_train)
y_pred_test = regr.predict(x_test)
y_pred_train = regr.predict(x_train)
```

```
from sklearn.metrics import mean_absolute_error, mean_squared_error, median_absolute_error,
r2_score
print('MSE train: {:.3f}, test: {:.3f}'.format(
      mean_squared_error(y_train, y_pred_train),
      mean_squared_error(y_test, y_pred_test)))
print('R^2 train: {:.3f}, test: {:.3f}'.format(
      r2_score(y_train, y_pred_train),
      r2_score(y_test, y_pred_test)))
from sklearn.preprocessing import PolynomialFeatures

polynomial_features= PolynomialFeatures(degree=2)

x_poly_train = polynomial_features.fit_transform(x_train)
x_poly_test = polynomial_features.fit_transform(x_test)

model = LinearRegression()
model.fit(x_poly_train, y_train)
y_poly_pred_test = model.predict(x_poly_test)
y_poly_pred_train = model.predict(x_poly_train)
X = train_df[['Нефть..м3']].values
y = train_df['Дебит.жидкости'].values
x_poly = polynomial_features.fit_transform(X)
model = LinearRegression()
model.fit(x_poly, y)
y_poly_pred = model.predict(x_poly)
plt.scatter(X, y, s=10)
import operator
sort_axis = operator.itemgetter(0)
sorted_zip = sorted(zip(X,y_poly_pred), key=sort_axis)
X, y_poly_pred = zip(*sorted_zip)
plt.plot(X, y_poly_pred, color='red')
plt.xlabel('Нефть,м3')
plt.ylabel('Дебит жидкости')
plt.show()
rmse = np.sqrt(mean_squared_error(y,y_poly_pred))
```

```
r2 = r2_score(y,y_poly_pred)
print('MSE train: {:.3f}, test: {:.3f}'.format(
      mean_squared_error(y_train, y_poly_pred_train),
      mean_squared_error(y_test, y_poly_pred_test)))
print('R^2 train: {:.3f}, test: {:.3f}'.format(
      r2_score(y_train, y_poly_pred_train),
      r2_score(y_test, y_poly_pred_test)))
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import train_test_split

Y = df['Нефть..м3'].values
X = df[['Попутный.газ..м3','Жидкость..м3','Добыча.растворенного.газа..м3',
'Дебит.попутного.газа..м3.сут',
     'Диаметр.экспл.колонны', 'Диаметр.НКТ', 'Глубина.верхних.дыр.перфорации',
'Удлинение', 'Производительность.ЭЦН',
     'Глубина.спуска', 'JD.факт', 'Удельный.коэффициент', 'Коэффициент.продуктивности']]
x_train, x_test, y_train, y_test = train_test_split(X, Y, test_size=0.3, random_state=0)
regr = RandomForestRegressor(max_depth=10, random_state=0)
regr.fit(x_train, y_train)
print(regr.feature_importances_)
y_pred_test = regr.predict(x_test)
y_pred_train = regr.predict(x_train)
from sklearn.metrics import mean_absolute_error, mean_squared_error, median_absolute_error,
r2_score
print('MSE train: {:.3f}, test: {:.3f}'.format(
      mean_squared_error(y_train, y_pred_train),
      mean_squared_error(y_test, y_pred_test)))
print('R^2 train: {:.3f}, test: {:.3f}'.format(
      r2_score(y_train, y_pred_train),
      r2_score(y_test, y_pred_test)))
```