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 Инженерная школа информационных технологий и робототехники
Направление подготовки 09.04.04 Программная инженерия
ООП/ОПОП Технологии больших данных
Отделение школы (НОЦ) Информационных технологий

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА МАГИСТРАНТА

Тема работы
Выявление перспективности нефтяных месторождений на основе построенных прогнозных моделей

УДК 303.094.6:622.24

Обучающийся

Группа	ФИО	Подпись	Дата
8ПМ1И	Еременко Марк Сергеевич		10.06.2023 г.

Руководитель ВКР

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ОИТ ИШИТР	Губин Евгений Иванович	к.ф.-м.н.		10.06.2023 г.

КОНСУЛЬТАНТЫ ПО РАЗДЕЛАМ:

По разделу «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение»

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Профессор ОСГН ШБИП ТПУ	Спицына Любовь Юрьевна	к.э.н.		

По разделу «Социальная ответственность»

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ООД ШБИП	Антоневич О. А.	к.б.н.		

ДОПУСТИТЬ К ЗАЩИТЕ:

Руководитель ООП, должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ОИТ ИШИТР	Губин Е. И.	к.ф.-м.н.		

Томск – 2023 г.

ПЛАНИРУЕМЫЕ РЕЗУЛЬТАТЫ ОСВОЕНИЯ ООП
по направлению 09.04.04 «Программная инженерия»

Код компетенции	Наименование компетенции
Универсальные компетенции	
УК(У)-1	Способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, вырабатывать стратегию действий
УК(У)-2	Способен управлять проектом на всех этапах его жизненного цикла
УК(У)-3	Способен организовывать и руководить работой команды, вырабатывая командную стратегию для достижения поставленной цели
УК(У)-4	Способен применять современные коммуникативные технологии, в том числе на иностранном (-ых) языке (-ах), для академического и профессионального взаимодействия
УК(У)-5	Способен анализировать и учитывать разнообразие культур в процессе межкультурного взаимодействия
УК(У)-6	Способен определять и реализовывать приоритеты собственной деятельности и способы ее совершенствования на основе самооценки
Общепрофессиональные компетенции	
ОПК(У)-1	Способен самостоятельно приобретать, развивать и применять математические, естественно-научные, социально-экономические и профессиональные знания для решения нестандартных задач, в том числе в новой или незнакомой среде и в междисциплинарном контексте
ОПК(У)-2	Способен разрабатывать оригинальные алгоритмы и программные средства, в том числе с использованием современных интеллектуальных технологий, для решения профессиональных задач
ОПК(У)-3	Способен анализировать профессиональную информацию, выделять в ней главное, структурировать, оформлять и представлять в виде аналитических обзоров с обоснованными выводами и рекомендациями
ОПК(У)-4	Способен применять на практике новые научные принципы и методы исследований
ОПК(У)-5	Способен разрабатывать и модернизировать программное и аппаратное обеспечение информационных и автоматизированных систем
ОПК(У)-6	Способен самостоятельно приобретать с помощью информационных технологий и использовать в практической деятельности новые знания

Код компетенции	Наименование компетенции
	и умения, в том числе в новых областях знаний, непосредственно не связанных со сферой деятельности
ОПК(У)-7	Способен применять при решении профессиональных задач методы и средства получения, хранения, переработки и трансляции информации посредством современных компьютерных технологий, в том числе, в глобальных компьютерных сетях
ОПК(У)-8	Способен осуществлять эффективное управление разработкой программных средств и проектов
Профессиональные компетенции	
ПК(У)-1	Способен к созданию вариантов архитектуры программного средства
ПК(У)-2	Способен разрабатывать и администрировать системы управления базам данных
ПК(У)-3	Способен управлять процессами и проектами по созданию (модификации) информационных ресурсов
ПК(У)-4	Способен проектировать и организовывать учебный процесс по образовательным программам с использованием современных образовательных технологий
ПК(У)-5	Способен осуществлять руководство разработкой комплексных проектов на всех стадиях и этапах выполнения работ

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 Инженерная школа информационных технологий и робототехники
Направление подготовки 09.04.04 Программная инженерия
ООП/ОПОП Технологии больших данных
Отделение школы (НОЦ) Информационных технологий

УТВЕРЖДАЮ:
Руководитель ООП
_____ _____ Губин Е. И.
(подпись) (дата) (Ф.И.О.)

ЗАДАНИЕ
на выполнение выпускной квалификационной работы

Обучающийся:

Группа	ФИО
8ПМ1И	Еременко Марк Сергеевич

Тема работы:

Выявление перспективности нефтяных месторождений на основе построения прогнозных моделей	
Утверждена приказом директора (дата, номер)	№ 146-39/с от 26.05.2023 г.

Срок сдачи обучающимся выполненной работы:	10.06.2023 г.

ТЕХНИЧЕСКОЕ ЗАДАНИЕ:

Исходные данные к работе *(наименование объекта исследования или проектирования; производительность или нагрузка; режим работы (непрерывный, периодический, циклический и т. д.); вид сырья или материал изделия; требования к продукту, изделию или процессу; особые требования к особенностям функционирования (эксплуатации) объекта или изделия в плане безопасности эксплуатации, влияния на окружающую среду, энергозатратам; экономический анализ и т. д.)*	Объектом исследования разработки являются данные, полученные при добыче нефтепродуктов методом гидравлического разрыва нефтеносного пласта.
Перечень разделов пояснительной записки, подлежащих исследованию, проектированию и разработке *(аналитический обзор по литературным источникам с целью выяснения достижений мировой науки техники в рассматриваемой области; постановка задачи исследования, проектирования, конструирования; содержание процедуры исследования, проектирования, конструирования; обсуждение результатов выполненной*	1. Обзор методов интеллектуального анализа данных 2. Методологии подготовки исходных данных 3. Рассмотрение модели логистической регрессии 4. Прогнозное моделирование

работы; наименование дополнительных разделов, подлежащих разработке; заключение по работе)	5. Работа над разделом по финансовому менеджменту, ресурсоэффективности и ресурсосбережения. 6. Работа над разделом по социальной ответственности.
Перечень графического материала *(с точным указанием обязательных чертежей)*	1. Диаграмма Ганта

Консультанты по разделам выпускной квалификационной работы
(с указанием разделов)

Раздел	Консультант
Основная часть	доцент ОИТ ИШИТР, к.ф.-м.н., Губин Е.И.
Финансовый менеджмент, ресурсоэффективность и ресурсосбережение	Профессор ОСГН ШБИП ТПУ, к.э.н., Спицына Л.Ю.
Социальная ответственность	доцент ООД ШБИП, к.б.н., Антоневич О. А.
Раздел на английском языке	Доцент ОИЯ ТПУ, к.филос.н, Уткина А.Н.

Названия разделов, которые должны быть написаны на иностранном языке:

References and acronyms
Introduction
Intellectual Data Analysis
Methodology and preparation of input data
References

Дата выдачи задания на выполнение выпускной квалификационной работы по линейному графику	01.03.2023 г.

Задание выдал руководитель ВКР:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ОИТ ИШИТР	Губин Евгений Иванович	к.ф.-м.н., доцент		01.03.2023 г.

Задание принял к исполнению обучающийся:

Группа	ФИО	Подпись	Дата
8ПМ1И	Еременко Марк Сергеевич		01.03.2023 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)

Инженерная школа Информационных технологий и робототехники
Направление подготовки (ООП / ОПОП) 09.04.04 Программная инженерия
Уровень образования магистратура
Отделение школы (НОЦ) Информационных технологий
Период выполнения весенний семестр 2022 /2023 учебного года

КАЛЕНДАРНЫЙ РЕЙТИНГ-ПЛАН
выполнения выпускной квалификационной работы

Обучающийся:

Группа	ФИО
8ПМ1И	Еременко Марк Сергеевич

Тема работы:

Выявление перспективности нефтяных месторождений на основе построения прогнозных моделей

Срок сдачи обучающимся выполненной работы:	10.06.2023 г.

Дата контроля	Название раздела (модуля) / вид работы (исследования)	Максимальный балл раздела (модуля)
10.06.2023	Основная часть	70
10.06.2023	Финансовый менеджмент, ресурсоэффективность и ресурсосбережение	10
10.06.2023	Социальная ответственность	10
10.06.2023	Раздел на английском языке	10

СОСТАВИЛ:
руководитель ВКР

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

СОГЛАСОВАНО:
руководитель ООП

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

Задание принял к исполнению обучающийся:

Группа	ФИО	Подпись	Дата
8ПМ1И	Еременко Марк Сергеевич		01.03.2023 г.

РЕФЕРАТ

Выпускная квалификационная работа 101 с., 14 рисунков, 21 табл., 40 источников, 3 прил.

Ключевые слова: анализ месторождений, логистическая регрессия, анализ данных, обработка данных, прогнозное моделирование, DATA MINING, моделирование.

Объектом исследования разработки являются данные, полученные при добыче нефтепродуктов методом гидравлического разрыва нефтеносного пласта.

Цель работы – построение прогнозной модели для оценки эффективности промышленной эксплуатации будущих нефтяных скважин.

В ходе работы проводился анализ методики обработки и анализа данных.

В результате разработана методология для выявления перспективных нефтяных месторождений с использованием прогнозных моделей, основанная на теории больших данных.

Область применения: разработка нефтяных месторождений

Содержание

Введение

Нефтегазовая промышленность играет ключевую роль в мировой экономике, обеспечивая основной источник энергии для промышленности, транспорта и домашнего потребления. Однако с постепенным истощением традиционных месторождений нефти и газа, поиск новых перспективных участков становится все более актуальным и критичным для обеспечения стабильного снабжения ресурсами.

Горючие полезные ископаемые, такие как нефть, являются сложной природной смесью углеводородов с примесями не углеводородных соединений. В зависимости от их состава, давления и температуры, углеводороды могут находиться в различных состояниях - твердом, жидком или газообразном. В некоторых условиях, часть углеводородов может находиться в жидком состоянии, тогда как другая часть будет в газообразном состоянии. Смеси углеводородов, которые находятся в жидком состоянии как в пластовых, так и в поверхностных условиях, называются нефтью. Состав нефти крайне сложен и разнообразен, и он может значительно изменяться даже в пределах одного месторождения. Физико-химические свойства нефти, включая ее товарные качества, определяются ее составом. Каждое нефтяное месторождение имеет уникальный состав и свойства нефти. В России эксплуатируется более 1300 нефтяных месторождений, а в мире их количество превышает 25 тысяч. Состав нефти может быть изменен в процессе добычи, при ее движении по пласту, в скважине, системах сбора и транспорта, а также при контакте с другими жидкостями и газами. [1].

В последние годы развитие технологий и научных исследований привело к возникновению новых методов и подходов в области геолого-геофизического моделирования и прогнозирования. Одним из таких подходов является построение прогнозных моделей в нефтегазовой отрасли.

Развитие информационных технологий, особенно в области обработки больших данных влечет за собой повышение требований к качеству исходных данных. Большинство реальных данных имеют слабую структуру, поэтому для

их грамотного анализа важно обеспечить их чистоту и надежность. Сбор и подготовка исходных данных являются одними из наиболее сложных и трудоемких этапов в анализе больших объемов данных, занимающих до 80% времени исследования. Однако использование статистических методов и современного программного обеспечения позволяет существенно сократить затраты времени и ресурсов на этот этап и повысить эффективность и качество конечных результатов.

Степень подготовки исходных данных имеет огромное влияние на предсказательную силу прогнозных моделей. При тщательной и корректной подготовке данных можно увеличить предсказательную точность традиционных моделей примерно на 20%. Это подчеркивает важность правильного подхода к сбору, очистке и форматированию данных. [2, 3].

Цель данной дипломной работы состоит в разработке методологии для выявления перспективных нефтяных месторождений, при добыче нефтепродуктов методом гидравлического разрыва нефтеносного пласта, с использованием прогнозных моделей, основанных на теории больших данных. В основе этой методологии лежит комплексный анализ различных факторов, включая геологические, геофизические, геохимические и инженерные данные, а также входные параметры скважин и исторические данные о производительности.

Данная работа имеет практическую значимость, поскольку позволяет применить систематический подход к разработке прогнозной модели в нефтегазовой отрасли, особенно это относится к методам гидравлического разрыва пласта. Результаты и выводы работы будут полезны для специалистов, занимающихся выбором наиболее перспективных месторождений.

Для достижения поставленной цели в работе будут использованы современные научные и технические источники, включая научные статьи, журналы и книги.

Определения, обозначения и сокращения

В данной работе применены следующие термины с соответствующими определениями:

целевая функция: вещественная или целочисленная функция нескольких переменных, подлежащая оптимизации (минимизации или максимизации) в целях решения некоторой оптимизационной задачи.

объясняющая переменная (Explanatory Variable): в экономико-статистических моделях (например, моделях регрессионного анализа) – то же, что независимая переменная, экзогенная переменная. Соответственно, зависимую переменную называют объясняемой.

мультиколлинеарность (multicollinearity): в эконометрике (регрессионный анализ) – 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.

Ниже представлен список используемых обозначений и сокращений:

DATA MINING – интеллектуальный анализ данных;

CRISP-DM – Cross-Industry Standard Process for Data Mining.

ROC – Receiver Operating Characteristic

AUC – Area Under Curve

1 Интеллектуальный анализ данных

Data Mining в соответствии с термином, введенном Г. Шапиро в 1989г., – это собирательное название, используемое для обозначения совокупности методов обнаружения в данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности. Data Mining можно дословно перевести как углубленный анализ данных. Сегодня компании используют Big Data для выявления скрытых закономерностей в исходных данных, построения прогнозных моделей, оптимизации операций, предотвращения угроз мошенничества и т.д. За последние два года такие компании, как IBM, Google, Amazon, Uber, создали сотни рабочих мест для программистов и специалистов Data Science. В основу методов Data Mining входят методы классификации, моделирования и прогнозирования, основанные на применении деревьев решений, искусственных нейронных сетей, генетических алгоритмов, эволюционного программирования, ассоциативной памяти, нечёткой логики, регрессионный анализ и т.п. Такие методы предполагают некоторые априорные представления об анализируемых данных [4].

В настоящее время Data Mining («интеллектуальный анализ данных») определяется как процесс поиска скрытых закономерностей в больших объемах данных, которые часто не структурированы и имеют разнообразные форматы (в виде чисел, текста, фото и т.п.). Большую часть времени интеллектуального анализа данных приходиться тратить на подготовку данных: очистку, агрегирование, преобразование и моделирование. Другой проблемой является то, что статистические модели часто строятся на данных с большим количеством наблюдений или переменных. Для масштабируемости необходимо тщательно выбирать и применять статистические методы. Как только у исследователей есть данные, они могут начать строить статистические модели. Из ряда возможных моделей в результате тестирования выбирается наилучшая. Часто бизнес-требования определяют

вид численной модели и во многих случаях это оказывается модель логистической регрессии, которая позволяет оценить вклад входящих переменных и сделать надежные прогнозные оценки [5].

1.1 Методология CRISP-DM

В современном информационном обществе данные играют все более важную роль в принятии управленческих решений и развитии бизнеса. Однако, собранные данные могут быть бесполезными, если не применять систематический и структурированный подход к их анализу и использованию. В этом контексте методология CRISP-DM представляет собой основу для разработки прогнозных моделей и извлечения ценной информации из данных [6].

CRISP-DM представляет собой универсальный стандартный процесс, разработанный для решения задач по обработке и анализу данных, включая построение прогнозных моделей. Методология включает шесть последовательных фаз, каждая из которых имеет свои специфические задачи и результаты (рисунок 1).

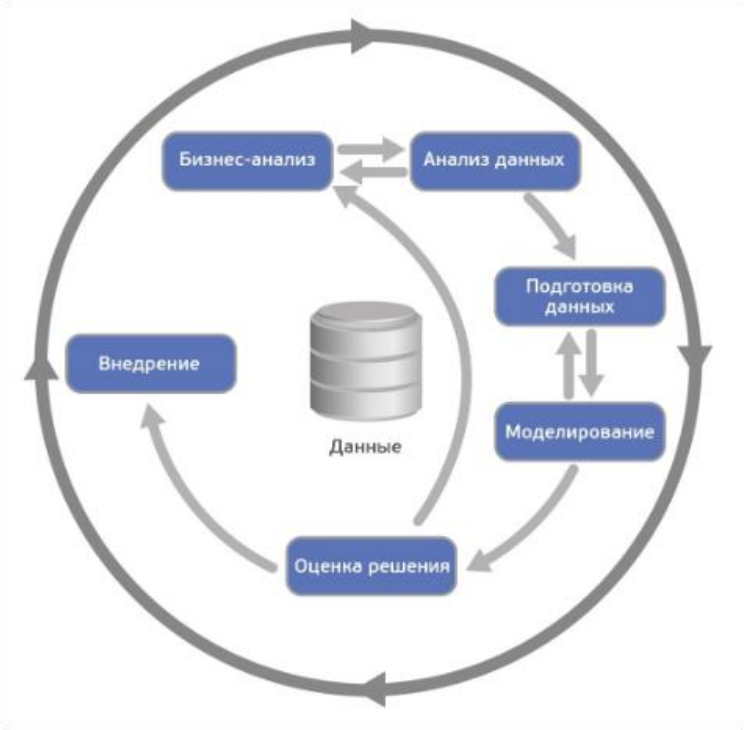

Рисунок 1 – Методология CRISP-DM

Эти фазы включают понимание бизнес-проблемы, понимание данных, подготовку данных, моделирование, оценку и внедрение результатов.

1. Фаза бизнес-анализа (Business understanding) – это первоначальный этап процесса, на котором определяются цели бизнеса и формулируются требования к ожидаемым результатам.
2. Фаза анализа данных (Data understanding) начинается собранием данных и включает их описание, изучение и проверку качества.
3. Фаза подготовки данных (Data preparation) включает в себя создание выборок, конструирование признаков, очистку, интеграцию и форматирование данных.
4. Фаза моделирования (Modeling) включает выбор, обучение и оценку качества моделей.
5. Фаза оценки результатов (Evaluation) включает оценку процесса анализа данных, полученных результатов и определение следующих шагов.
6. Фаза внедрения (Deployment) предполагает внедрение разработанной модели, ее мониторинг и получение обратной связи [7].

Преимущество CRISP-DM заключается в гибкости и применимости в различных отраслях и сферах деятельности. Методология обеспечивает систематический подход к работе с данными, позволяет эффективно управлять проектами по анализу данных и минимизировать риски. Благодаря этому, она получила широкое распространение и признание в индустрии и научном сообществе [8].

В процессе работы будут рассмотрены фазы CRISP-DM и их применение на практике. Также будут рассмотрены современные подходы и инструменты, связанные с каждой фазой методологии.

1.2 Методология SEMMA

Институт SAS определяет интеллектуальный анализ данных как процесс выборки, изучения, изменения, моделирования и оценки (SEMMA) больших объемов данных для выявления ранее неизвестных шаблонов, которые могут

быть использованы в качестве бизнес-преимущества. Процесс интеллектуального анализа данных применим в различных отраслях промышленности и предоставляет методологии для таких разнообразных бизнес-проблем, как обнаружение мошенничества, удержание клиентов, маркетинговый анализ баз данных, сегментация рынка, анализ рисков, анализ сходства, удовлетворенность клиентов, прогноз банкротства и анализ кредитного портфеля [9].

На рисунке 2 показан известный логический алгоритм SEMMA.

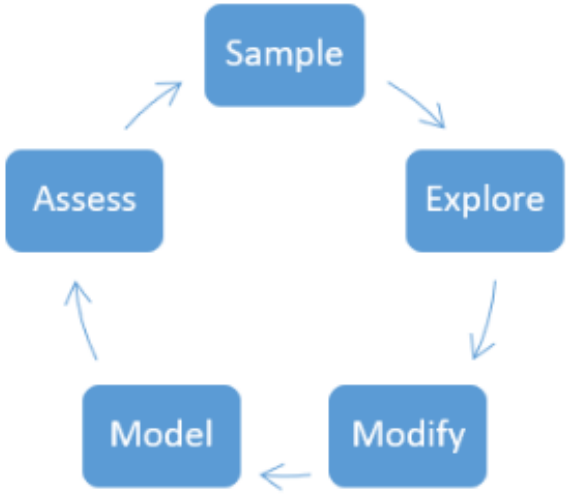

Рисунок 2 – Методология SEEMA

Каждый этап методологии SEMMA выполняет определенные функции, обеспечивая целостность процесса интеллектуального анализа данных.

Первый этап, SAMPLE (выборка), включает формирование начального набора данных для моделирования. Набор данных должен быть достаточно большим, чтобы содержать необходимую информацию, и в то же время ограниченным, чтобы обеспечить эффективное использование.

На этапе EXPLORE (исследование) происходит выявление ассоциаций и проведение визуального и интерактивного статистического анализа данных. Цель здесь - понять данные путем обнаружения связей между переменными и выявления как ожидаемых, так и непредвиденных взаимосвязей с помощью визуализации.

Этап MODIFY (изменение) включает подготовку данных для анализа. Здесь создаются дополнительные переменные или модифицируются существующие, чтобы обеспечить анализ. Также проводится оценка выбросов и отсутствующих данных, изменение способа использования входных переменных и проведение различных анализов, таких как кластерный анализ или сети Кохрена.

На этапе MODEL (моделирование) применяются методы построения и обработки моделей интеллектуального анализа данных, такие как искусственные нейронные сети, деревья принятия решений, регрессионный анализ и т.д.

ASSESS (оценка) включает сравнение результатов моделирования с планируемыми показателями, анализ надежности и полезности созданных моделей.

Методология SEMMA предоставляет набор инструментов для каждой из указанных категорий и облегчает процесс интеллектуального анализа данных. Она скорее является сводом рекомендаций, чем жесткими правилами.

2 Методология подготовки исходных данных

Сбор и подготовка исходных данных – это один из самых сложных и трудоемких этапов в процессе анализа большого объема информации, который может занимать до 80% рабочего времени специалиста. Современные статистические методики и современное программное обеспечение позволяют значительно сократить время, финансовые и временные затраты на данный этап, а также повысить качество конечного результата [2].

В процессе предварительной обработки исходных данных выполняются следующие операции: очистка данных от ошибок и аномалий, исследование выбросов и повторяющихся записей. Важной задачей этого этапа является выявление наличия мультиколлинеарности между объясняющими переменными и при необходимости исключения или замены этих переменных. Процесс масштабирования позволяет привести исходные данные к единому цифровому формату, что существенно повышает точность прогнозных моделей [3].

2.1 Проверка исходных данных на ошибки

Проверка исходных данных на ошибки или опечатки заключается в анализе и оценке данных с целью выявления потенциальных ошибок при их записи. Эта проверка направлена на обнаружение случайных или некорректных значений, а также на выявление возможных опечаток, которые могут исказить исходные данные и повлиять на результаты анализа.

В процессе проверки данных на ошибки можно применять различные методы и подходы. Один из распространенных методов – это анализ статистических показателей (описательной статистики), таких как среднее значение, медиана, стандартное отклонение и т.д. [10]. Некорректные значения могут быть обнаружены, если они существенно отличаются от ожидаемых статистических показателей. Также можно использовать методы визуализации данных, например, построение графиков или диаграмм, чтобы визуально выявить аномалии или необычные значения. Например, если

данные представлены в виде временных рядов, то графическое отображение может помочь выявить выбросы или неправдоподобные значения [11].

Другой подход – это сравнение данных с известными правилами или ожидаемыми значениями. Если данные не соответствуют заранее определенным правилам или находятся за пределами допустимых диапазонов, то это может указывать на наличие ошибок или опечаток.

В целом, проверка исходных данных на ошибки требует внимательного анализа и экспертного подхода. Результаты этой проверки могут послужить основой для дальнейшей предобработки данных и обеспечения их надежности и качества перед построением прогнозных моделей.

2.2 Проверка исходных данных на отсутствующие значения

Идентификация и обработка пропущенных данных являются важными этапами в анализе данных. Пропущенные значения могут возникать по разным причинам, таким как ошибки при сборе данных, технические проблемы или отсутствие должной информации. Однако, наличие отсутствующих значений может серьезно исказить результаты анализа и прогнозирования. В начале процесса проверки на отсутствие данных, мы осуществляем идентификацию пропущенных значений в различных переменных и столбцах набора данных. Для этого проверяем каждую запись данных и выявляем пропущенные или некорректные значения.

После идентификации пропущенных данных, следующим шагом является выбор оптимальной стратегии их обработки. Существует несколько методов обработки пропущенных данных, включая удаление записей с пропущенными значениями, замену пропущенных значений средними или медианными значениями, использование интерполяции, статистических моделей и других подходов [3, 12].

Конкретная стратегия обработки пропущенных данных зависит от контекста и особенностей набора данных. Учитываются тип переменной

(категориальная или числовая), распределение данных, объем пропущенных значений, а также возможные причины и другие факторы [13].

Целью проверки и обработки пропущенных данных является обеспечение полноты и качества данных, чтобы исключить искажения в анализе и моделировании. Это позволяет получить более точные и достоверные результаты исследования и прогнозирования на основе доступных данных.

2.3 Проверка исходных данных на выбросы

Выбросы в данных – это аномальные значения, выделяющиеся из общей выборки. Проверка исходных данных на выбросы представляет собой процесс выявления и анализа наблюдений, которые значительно отклоняются от ожидаемого поведения или распределения данных. Выбросы могут возникать по разным причинам, таким как ошибки в сборе данных, аномальные события или естественные вариации в данных.

Одним из методов проверки на выбросы является анализ статистических показателей, таких как среднее значение, медиана и стандартное отклонение. Выбросы могут быть выявлены, если значения существенно отличаются от ожидаемых статистических показателей или находятся вне заданных границ [14].

Для обнаружения выбросов в данных также можно применять графические методы. В случае категориальных переменных гистограммы могут быть полезными инструментами. При работе с числовыми переменными, простейший подход к определению выбросов заключается в выделении всех наблюдений, которые не попадают в заданные квантили. Этот подход может быть визуализирован с помощью диаграмм размаха, также известных как "ящики с усами". Графическое представление данных в виде ящиков с усами позволяет наглядно выявить потенциальные выбросы и их распределение.

Использование графических методов, таких как гистограммы и диаграммы размаха, обеспечивает наглядность и интуитивное понимание данных, помогая выявить аномальные значения и потенциальные выбросы. [14].

Проверка на выбросы позволяет выявить потенциально ошибочные или аномальные данные, которые могут исказить результаты анализа. Если количество выбросов невелико, их можно удалить из анализа или заменить средним или модой, чтобы устранить их влияние. Однако, при большом количестве выбросов целесообразно выделить их в отдельную выборку для проведения дополнительного анализа, поскольку это может свидетельствовать о появлении нового феномена в данных или особенностях их распределения. Для борьбы с выбросами в числовых переменных могут применяться различные методы преобразования данных. Например, можно использовать «min-max нормализацию», которая масштабирует значения переменной в определенном диапазоне, чтобы сгладить экстремальные значения и сделать данные более сопоставимыми.

Применение соответствующих преобразований и методов обработки выбросов помогает справиться с их влиянием на анализ и достичь более надежных результатов.

2.4 Проверка исходных данных на наличие дублирующих строк

Проверка исходных данных на наличие дублирующих строк представляет собой процесс выявления и анализа повторяющихся записей в наборе данных. Дублирующие строки могут возникать по разным причинам, таким как технические ошибки при сборе или записи данных, повторные вводы информации или другие случайные события. В процессе проверки на наличие дублирующих строк, первоначально осуществляется идентификация повторяющихся записей в наборе данных. Это может быть выполнено путем сравнения значений различных полей или столбцов в каждой записи данных и выявления идентичных или очень похожих значений [2].

После идентификации дублирующих строк, следующим шагом является удаление дубликатов.

2.5 Проверка объясняющих переменных (атрибутов) на мультиколлинеарность

Проверка атрибутов на мультиколлинеарность – это процесс выявления и оценки степени линейной зависимости между различными объясняющими переменными в наборе данных. Мультиколлинеарность может возникать, когда две или более переменные сильно коррелируют друг с другом, что может привести к не единственности результатов прогнозных моделей.

Основная цель проверки на мультиколлинеарность состоит в определении, насколько сильно объясняющие переменные связаны друг с другом. Это может быть достигнуто путем вычисления корреляционных коэффициентов между переменными, таких как коэффициент Пирсона или коэффициент Спирмена. Значения корреляции близкие к 1 или -1 указывают на сильную линейную связь между переменными. Если обнаруживается высокая корреляция между объясняющими переменными, возникает необходимость принять меры. Возможными методами обработки мультиколлинеарности являются исключение одной из коррелирующих переменных, объединение переменных в новую, комбинированную переменную или применение методов регуляризации, таких как ридж-регрессия или лассо-регрессия [15].

Важно отметить, что проверка на мультиколлинеарность должна быть выполнена перед построением прогнозной модели, чтобы избежать проблем с неустойчивостью оценок параметров модели и недостоверными выводами. Эта проверка помогает улучшить качество моделирования и уменьшить возможность ошибочных интерпретаций результатов.

2.6 Выбор целевой переменной

Выбор целевой переменной является важным этапом при построении прогнозных моделей. Целевая переменная, также известная как зависимая переменная или переменная отклика, является основным объектом прогнозирования и определяет цель исследования. При выборе целевой переменной необходимо учитывать цели и задачи исследования, а также доступные данные. Целевая переменная должна быть тщательно определена, чтобы отражать интересующую нас характеристику или явление, которое мы хотим прогнозировать или объяснить. При выборе целевой переменной также важно учитывать ее доступность и качество данных. Целевая переменная должна быть измеримой и надежной, чтобы обеспечить точность и надежность прогнозов [3].

Процесс выбора целевой переменной может зависеть от типа прогнозной модели и контекста задачи. Например, в задачах прогнозирования временных рядов целевой переменной может быть будущее значение временного ряда. В задачах классификации целевая переменная может принимать категориальные значения и представлять собой классы или категории, которые необходимо предсказать.

Помимо этого, выбор целевой переменной может быть основан на предварительном исследовании и экспертном знании предметной области. Иногда важно учесть факторы, которые могут влиять на целевую переменную и включить их в модель в качестве объясняющих переменных.

В целом, выбор целевой переменной требует внимательного анализа целей исследования, доступности и качества данных, а также экспертного знания предметной области. Это поможет построить прогнозные модели, которые эффективно прогнозируют интересующие нас явления и вносят практическую пользу.

2.7 Разбиение исходных данных на тренировочную и тестовую выборки

Для обеспечения адекватности модели и проверки ее точности на исходных исторических данных рекомендуется разделить данные на две или три независимые выборки.

Если планируется обучение модели на небольшом количестве исходных данных, то наиболее распространенным подходом является разделение на две выборки: тренировочную и тестовую. В процессе обучения моделей на тренировочной выборке осуществляется оптимизация их параметров, а на тестовой выборке проводится оценка качества модели. Обычно соотношение числа наблюдений в тренировочной и тестовой выборках составляет примерно 70-80% и 30-20% соответственно. Это соотношение определяется доступным объемом исходных данных. Важно отметить, что модель, обученная на небольшом количестве тестовых данных, может иметь большую дисперсию, то есть ее результаты на различных выборках могут значительно отличаться. При наличии среднего объема выборки (тысячи или десятки тысяч наблюдений) распространенным подходом является использование соотношения 70% к 30% или 80% к 20% между тренировочной и тестовой выборками.

Применение правильного разбиения выборок позволяет оценить и улучшить качество модели, а также оценивать ее производительность на новых, неизвестных данных [16].

При обучении моделей с большим объемом исходных данных рекомендуется применять разбиение на три выборки: тренировочную, валидационную и тестовую. Это позволяет решить проблему переобучения и получить более надежную оценку модели. На валидационной выборке происходит сравнение результатов нескольких моделей и выбор наилучшей. Стандартное соотношение размеров трех выборок для средних объемов данных составляет 60%, 20% и 20%. Рекомендации относительно соотношения размеров двух выборок также применимы в этом случае.

При разделении исходных данных на выборки возникает проблема репрезентативности. Все выборки должны иметь аналогичное соотношение классов целевой переменной, как и в исходной выборке. Это гарантирует, что каждая выборка будет достаточно представительной для обучения и оценки модели.

Вышеперечисленные схемы разбиения данных представлены на рисунке 3. Применение этих схем помогает обеспечить адекватное обучение моделей и достоверную оценку их производительности.

Рисунок 3 – Разбиение исходных данных

3 Логистическая регрессия

Модель логистической регрессии – это статистическая модель, которая используется для прогнозирования вероятности возникновения определенного события или классификации наблюдений на основе набора предикторов или объясняющих переменных. Она является одной из основных моделей в области машинного обучения и статистики, широко применяемой в различных областях, включая медицину, маркетинг, финансы и социальные науки.

Основной смысл модели логистической регрессии заключается в анализе и прогнозировании бинарных или категориальных зависимых переменных. Она позволяет определить вероятность принадлежности наблюдения к определенному классу на основе значения предикторов. В отличие от модели линейной регрессии, которая предсказывает размерную, непрерывную зависимую переменную, логистическая регрессия использует логистическую функцию для использования вероятностного подхода при прогнозном анализе [17].

Преимущество модели логистической регрессии заключается в ее интерпретируемости. Она позволяет оценить вклад каждого предиктора и определить их статистическую значимость в объяснении зависимой переменной. Коэффициенты регрессии в модели логистической регрессии интерпретируются как логарифмы отношения шансов, что позволяет сделать выводы о влиянии каждого предиктора на вероятность принадлежности к определенному классу. Кроме того, модель логистической регрессии обладает хорошей способностью к обобщению на новые данные и хорошо справляется с задачами классификации, особенно в случаях, когда классы являются линейно разделимыми или близкими к линейно разделимым [17,18].

Однако следует учитывать ограничения модели логистической регрессии. Она предполагает линейную зависимость между предикторами и логарифмом шансов, а также независимость наблюдений. В случае наличия

нелинейных зависимостей или нарушения предпосылок модели, может потребоваться применение альтернативных моделей [18].

Логистическая регрессия работает очень похоже на линейную регрессию, но с биномиальной переменной отклика. Самым большим преимуществом является тот факт, что вы можете использовать непрерывные объясняющие переменные и легче обрабатывать более двух объясняющих переменных одновременно. Несмотря на кажущуюся тривиальность, эта последняя характеристика важна, когда нас интересует влияние различных объясняющих переменных на переменную отклика. Если мы рассматриваем несколько объясняющих переменных независимо, мы игнорируем ковариацию между переменными и подвергаемся смешивающим эффектам, как было продемонстрировано в приведенном выше примере, когда влияние лечения на вероятность смерти было частично скрыто эффектом возраста. Логистическая регрессия смоделирует вероятность результата на основе индивидуальных характеристик [18]. Поскольку вероятность – это отношение, то, что на самом деле будет смоделировано, – это логарифм вероятности, заданный следующим образом:

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m \quad (1)$$

где π – указывает на вероятность события;

β_0 – коэффициенты регрессии, связанные с контрольной группой;

x_i – объясняющие переменные.

4 Прогнозное моделирование

4.1 Работа с данными

На первом этапе работы были получены необходимые для анализа данные из исходного дата сета. Дата сет представляет из себя набор экспериментальных показателей скважин, имеющих трещиноватую структуру грунта. Необходимо оценить важность и влияние измеренных входных параметров на величину добычи нефти. В дальнейшем построить прогнозную модель для оценки величины добычи нефти от входных параметров и тем самым оценить эффективность и целесообразность будущей добычи.

В качестве входных параметров (переменных) были взяты следующие исходные данные (Таблица 1).

Таблица 1 – Список исходных переменных и их размерность

Имя переменной	Расшифровка
R	Радиус зоны дренирования, **м**
Qgrp	Количество стадий ГРП, **шт**
W	Мощность продуктивного пласта, **м**
Pplast	Пластовое давление, **атм**
K	Проницаемость, **мД**
Noil	Вязкость нефти, **сП**
Azimut	Азимут распространения трещины, **градусы**
Mp	Масса проппанта, **тонн**
St	Полудлина трещины (длина одного крыла), **м**
Ht	Высота трещины, **м**
Lt	Ширина трещины, **мм**
Pz	Забойное давление, **атм**
DetltaP	Депрессия (разница пластового и забойного давлений), **атм**
Df	Дебит жидкости скважины, **м3/сут**
Doil	Дебит нефти скважины, **м3/сут**
Dw	Дебит воды скважины, **м3/сут**

Целевой переменной была выбрана Doil, которая определяет дебит нефти скважины ($м^3$/сут). На рисунке 4 показана величина добычи Doil в зависимости от номера скважины.

Рисунок 4 – Дебит нефти скважины

Из экспертных оценок можно предположить, что наиболее эффективная добыча нефти находится выше среднего значения, равного 35 $м^3$/сут. Будем использовать это при построении бинарной целевой функции «Y».

«Y» принимает значение 0, если Дебит нефти скважины меньше среднего значения, и равен 1, если Дебит нефти скважины, больше или равен среднему значению (35 $м^3$/сут).

Перед тем как строить модель по исходным данных, необходимо провести их качественный и количественный анализ. Эффективность и предсказательная сила будущей модели во многом зависит от качества исходных данных. На основании методологии подготовки исходных, описанной в начале работы был проведен анализ данных.

Для решения поставленной задачи в качестве технического инструмента был использован язык программирования Python. Полный листинг программного кода приведен в Приложении Г.

НА первом этапе были определенны типы данных. Из рисунка 5 видно, что исходные данные носят числовой характер.

```
R           int64
Qgrp        int64
W         float64
Pplast      int64
K         float64
Noil      float64
Azimut    float64
Mp        float64
St        float64
Ht        float64
Lt        float64
Pz          int64
DeltaP      int64
Df        float64
Doil      float64
Dw        float64
dtype: object
```

Рисунок 5 – Типы данных

На следующем этапе проверялись отсутствующие значения в данных. Таких значений в дата сете не оказалось (Рисунок 6).

```
R         0
Qgrp      0
W         0
Pplast    0
K         0
Noil      0
Azimut    0
Mp        0
St        0
Ht        0
Lt        0
Pz        0
DeltaP    0
Df        0
Doil      0
Dw        0
dtype: int64
```

Рисунок 6 – Проверка данных на пропущенные значения

Далее данные проверялись на выбросы с помощью построения диаграммы размаха. Из рисунка 7 видно, что по некоторым величинам имеются выбросы.

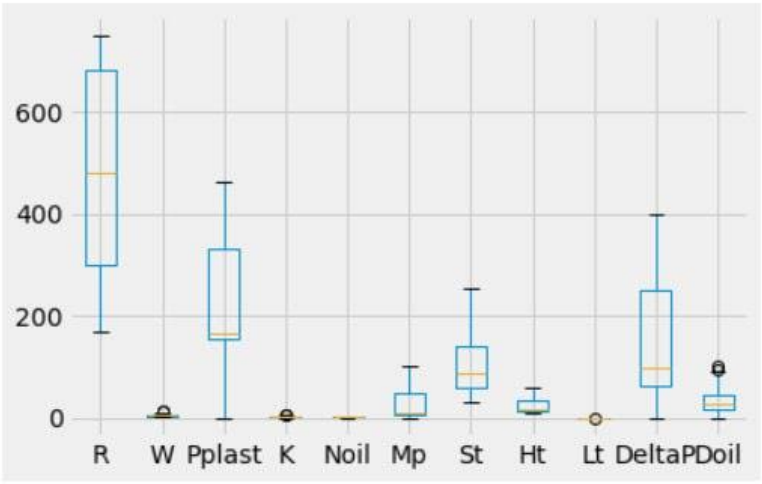

Рисунок 7 – Диаграмма размаха

Так как модель логистическое регрессии чувствительна к выбросам необходимо отчистить дата сет от их наличия. В соответствие с изложенной выше методологией выбросы были заменены на средние значения.

На следующем этапе данные проверялись на мультиколлинеарность с помощью построения корреляционной матрицы (рисунок 8).

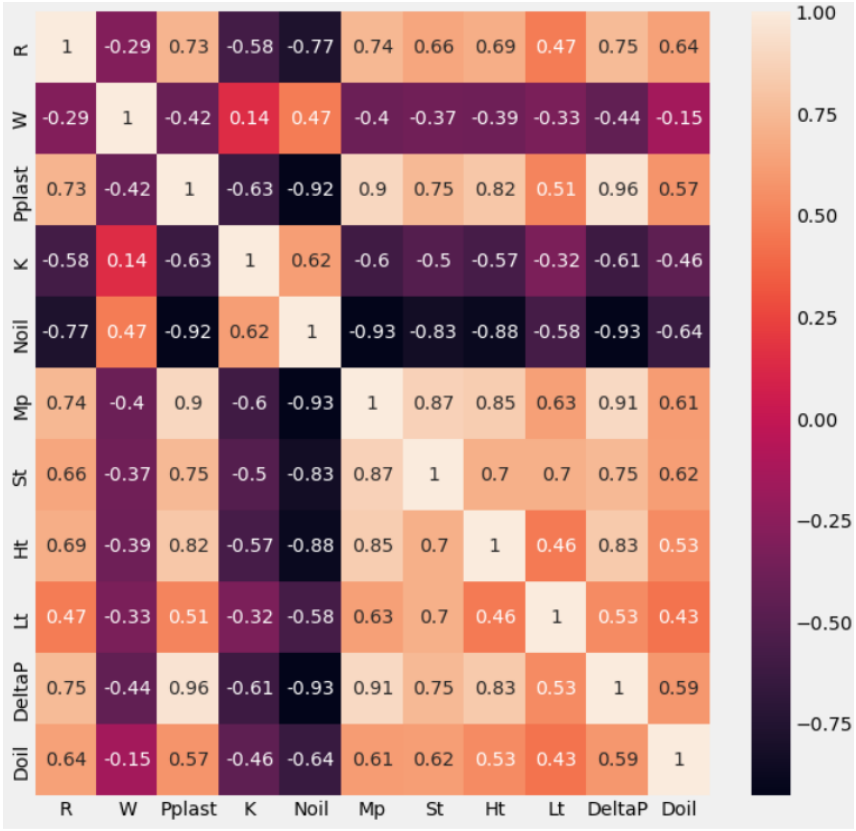

Рисунок 8 – Корреляционная матрица

С учетом вышеприведенной методологии сильно коррелируемые значения исходных переменных были исключены из анализа, что видно из рисунка 9, где количество атрибутов сократилось на 5 переменных.

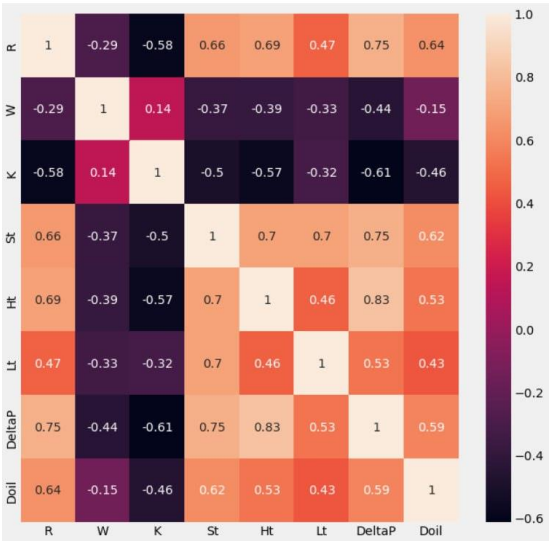

Рисунок 9 – Корреляционная матрица без исключенный переменных

Исходя из количества исходных данных (185 наблюдений) используется только тренировочная и тестовая выборки в соотношении 139 на 49 наблюдений (75% к 25% соответственно).

В итоге проведенных процедур по подготовки исходных данных для использования логистической регрессии для оценки вероятность события (p) от входных параметров (x), мы получили 7 входных (вместо 13) безразмерных переменных и целевую переменную «Y».

4.2 Моделирование и анализ логистической модели

На следующем этапе переходим непосредственно к построению модели логистической регрессии. На рисунке 10 представлен исходный код построения логистической модели на языке Python с использованием библиотеки Scikit-learn.

```
x = df.loc[:, ('R','W','K','St','Ht','Lt','DeltaP')]
y = df.loc[:, 'Doil']
X_train, X_test, y_train, y_test = train_test_split(x, y,
                                                    train_size=0.75,
                                                    random_state=42)
model = LogisticRegression().fit(X_train, y_train)
```

Рисунок 10 – Исходный код построения логистической модели на языке Phyton

Результатом проведенного анализа была получена регрессионная модель вероятности эффективности работы скважин «Y» нефтяного месторождения в зависимости от входных параметров данного месторождения R (Радиус зоны дренирования), W (Мощность продуктивного пласта), K (Проницаемость), St (Полудлина трещины), Ht (Высота трещины), Lt (Ширина трещины), DetltaP (разница пластового и забойного давлений),

На рисунке 11 показана ROC кривая тестовой выборки прогнозной модели, величина которой равняется AUC, что говорит о хорошей прогнозной силе построенной модели.

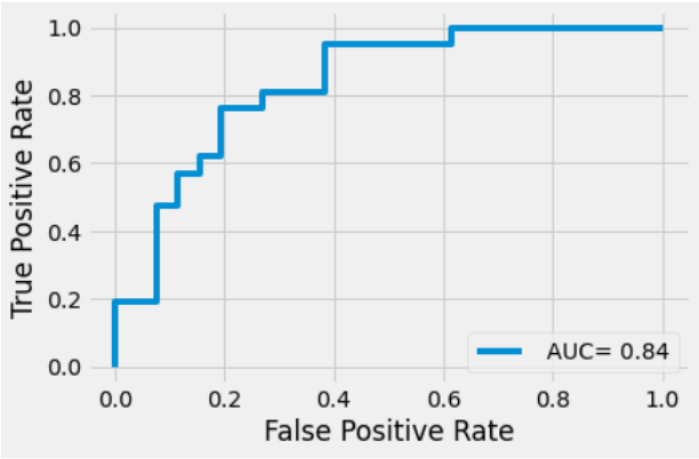

Рисунок 11 – ROC кривая тестовой выборки прогнозной модели

Подтверждением этого служит матрица ошибок («confusion matrix»), представленная на рисунке 12, из которой видно, что верный прогноз эффективных скважин находится в пределах 75%.

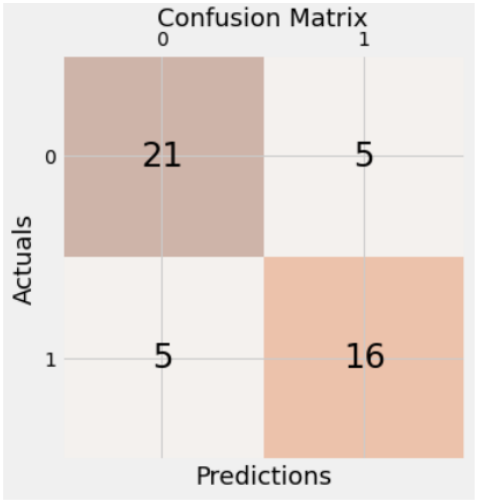

Рисунок 12 – Матрица ошибок

Хорошую точность модели также показывают приведённые метрики оценки (рисунок 13).

```
print('Precision: %.3f' % precision_score(y_test, y_pred))
print('Recall: %.3f' % recall_score(y_test, y_pred))
print('Accuracy: %.3f' % accuracy_score(y_test, y_pred))
```

```
Precision: 0.762
Recall: 0.762
Accuracy: 0.787
```

Рисунок 13 – Метрики

5 Финансовый менеджмент, ресурсоэффективность и ресурсосбережение

ЗАДАНИЕ ДЛЯ РАЗДЕЛА
«ФИНАНСОВЫЙ МЕНЕДЖМЕНТ, РЕСУРСОЭФФЕКТИВНОСТЬ И РЕСУРСОСБЕРЕЖЕНИЕ»

Обещающемуся:

Группа	ФИО
8ПМ1И	Еременко Марку Сергеевичу

Школа	ИШИТР	Отделение школы (НОЦ)	Отделение информационных технологий
Уровень образования	Магистратура	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	*Бюджет проекта – не более 374 361, 00 руб., в т.ч. затраты по оплате труда – не более 138 951, 00 руб.*
2. Нормы и нормативы расходования ресурсов	*Значение показателя интегральной ресурсоэффективности – не менее 4,7 баллов из 5,00*
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	*Проведение предпроектного анализа. Определение целевого рынка и проведение его сегментирования. Выполнение SWOT-анализа проекта*
2. Разработка устава научно-технического проекта	*Определение целей и ожиданий, требований проекта. Определение заинтересованных сторон и их ожиданий*
3. Планирование процесса управления НТИ: структура и график проведения, бюджет, риски и организация закупок	*Составление календарного плана проекта. Определение бюджета НТИ*
4. Определение ресурсной, финансовой, экономической эффективности	*Проведение оценки экономической эффективности разработки*
Перечень графического материала *(с точным указанием обязательных чертежей):*	
1. Сегментирование рынка *2. Анализ конкурентоспособности технического решения* *3. Матрица SWOT* *4. Инициация проекта* *5. Перечень работ* *6. График проведения НТИ* *7. Бюджет НТИ* *8. Оценка ресурсной, финансовой и экономической эффективности НТИ*	

Дата выдачи задания для раздела по линейному графику	

Задание выдал консультант:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Профессор ОСГН ШБИП ТПУ	Спицына Любовь Юрьевна	к.э.н.		

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
8ПМ1И	Еременко Марк Сергеевич		

5.1 Предпроектный анализ

5.1.1 Потенциальные потребители результатов исследования

Разработанную прогнозную модель в дальнейшем планируется применять в нефтегазовой отрасли для оценки эффективности скважин от входных параметров на основе теории больших данных. Также такая аналитика может быть использована для изучения оценки влияния экспериментальных входных параметров на дебит нефти.

Основные заинтересованные в продукте лица – крупные и средние компании, имеющие большое количество месторождений, требующих мониторинга и детального анализа отдельных пластов.

Сегментировать рынок услуг можно по размеру компании-заказчика и сфере применения разработки. Результат сегментирования представлен в таблице 2.

Таблица 2 – Карта сегментирования рынка

		Сфера применения разработки			
		Прогнозирование дебита скважин	Оптимизация работы скважин	Диагностика и прогнозирование неисправностей	Прогнозирование объемов запасов
Размер компании	Крупные				
	Средние				
	Мелкие				

	Фирма А		Фирма Б

Под фирмой А подразумевается множество фирм, представляющих собой корпоративные научно-исследовательские проектные институты, выполняющие функции проектного офиса. Под фирмой Б рассматриваются добывающие общества. Таким образом, основными заказчиками, на которые ориентирован программный продукт, являются крупные и средние компании, имеющие потребность в применения прогнозных моделей в нефтегазовой отрасли. Это помогает компаниям принимать обоснованные решения на основе анализа больших объемов данных и повышать эффективность операций добычи.

5.1.2 Анализ конкурентоспособности технического решения

Детальный анализ конкурирующих разработок, существующих на рынке, необходимо проводить систематически, поскольку рынки пребывают в постоянном движении. Такой анализ помогает вносить коррективы в научное исследование, чтобы успешнее противостоять своим соперникам. Важно реалистично оценивать сильные и слабые стороны разработок конкурентов.

Анализ конкурентоспособности технического решения был проведен с помощью оценочной карты. В основе данной технологии лежит нахождение средневзвешенного значения показателя качества и перспективности научной разработки. Позиция разработки и конкурентов оценивается по каждому показателю экспертным путем по пятибалльной шкале, где 1 – наиболее слабая позиция, а 5 – наиболее сильная. Веса показателей, определяемые экспертным путем, в сумме должны составлять 1. Анализ конкурентных технических решений определяется по формуле:

$$K = \sum B_i \cdot Б_i, \qquad (2)$$

где K – конкурентоспособность научной разработки или конкурента;

B_i– вес показателя (в долях единицы);

$Б_i$– балл i-го показателя.

Оценочная карта анализа для сравнения конкурентных технических решений представлена в таблице 3.

Таблица 3 – Оценочная карта для сравнения конкурентных технических решений (разработок)

Критерии оценки	**Вес крите-рия**	**Баллы**			**Конкурентоспо-собность**		
		$Б_ф$	$Б_{к1}$	$Б_{к2}$	$К_ф$	$К_{к1}$	$К_{к2}$
1	2	3	4	5	6	7	8
Технические критерии оценки ресурсоэффективности							
1.Помехоустойчивость	0,1	3	4	3	0,3	0,4	0,3
2.Экологичность	0,18	5	4	3	0,9	0,72	0,54
3.Надежность	0,05	4	4	3	0,2	0,2	0,15
4.Простота эксплуатации	0,1	5	3	3	0,5	0,3	0,3
5.Экономичность	0,09	5	3	3	0,45	0,27	0,27
Экономические критерии оценки эффективности							
1.Конкурентоспособность продукта	0,07	4	4	3	0,28	0,28	0,21
2.Уровень проникновения на рынок	0,07	3	5	5	0,21	0,35	0,35
3.Цена	0,07	5	2	3	0,35	0,14	0,21
4.Предполагаемый срок эксплуатации	0,08	5	5	5	0,4	0,4	0,4
5.Послепродажное обслуживание	0,06	4	5	4	0,24	0,3	0,24
6.Финансирование научной разработки	0,03	3	5	4	0,09	0,15	0,12
7.Срок выхода на рынок	0,04	5	4	4	0,2	0,16	0,16
8.Наличие сертификации разработки	0,06	3	5	4	0,18	0,3	0,24
Итого	**1**	**54**	**53**	**47**	**4,3**	**3,97**	**3,49**

Из оценочной карты можно сделать вывод о том, что разрабатываемая система является перспективной. Также у разрабатываемого решения наиболее важные критерии имеют высокие показатели, и суммарная оценка составляет 4,3.

5.1.3 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. Он проводится в несколько этапов.

Первый этап заключается в описании сильных и слабых сторон проекта, в выявлении возможностей и угроз для реализации проекта, которые проявились или могут появиться в его внешней среде (таблица 4).

Таблица 4 – Результаты первого этапа SWOT-анализа

	Сильные стороны научно-исследовательского проекта:	**Слабые стороны научно-исследовательского проекта:**
	Си1. Применение прогнозной модели позволяет улучшить стратегическое планирование в нефтегазовой отрасли Си2. Применение прогнозной модели на основе больших данных позволяет оптимизировать процессы и принимать более информированные решения Си3.Небольшая цена разработки, по сравнению с конкурентами	Сл1. Сильная зависимости результатов оценки от качества входных данных Сл2. Небольшой размер входной выборки Сл3. Результат моделирования носит вероятностный характер

Продолжение таблицы 4

Возможности: В1. Улучшение точности и надежности В2. Увеличение эффективности операций В3. Появление дополнительного спроса на программный продукт		
Угрозы: У1. Уход от использования нефтепродуктов У2. Появление более качественных аналогов модели У3. Появление более дешёвых аналогов		

В рамках второго этапа на основе предыдущих результатов были выявлены соответствия сильных и слабых сторон проекта внешних условиям окружающей среды. Интерактивная матрица проекта представлена в таблице 5.

Таблица 5 – Результаты второго этапа SWOT-анализа

Сильные стороны проекта				
Возможности проекта		С1	С2	С3
	В1	+	+	+
	В2	+	+	+
	В3	+	+	–
Слабые стороны проекта				
Возможности проекта		Сл.1	Сл.2	Сл. 3
	В1	+	+	+
	В2	+	+	–
	В3	–	+	–

Продолжение таблицы 5

Сильные стороны проекта				
Угрозы проекта		С1	С2	С3
	У1	0	+	–
	У2	0	0	–
	У3	+	0	0
Слабые стороны проекта				
Угрозы проекта		Сл.1	Сл.2	Сл.3
	У1	–	0	–
	У2	–	–	–
	У3	–	–	–

Составленная в ходе выполнения третьего этапа SWOT-анализа итоговая матрица представлена в таблице 6.

Таблица 6 – Итоговая матрица SWOT-анализа

	Сильные стороны научно-исследовательского проекта:	**Слабые стороны научно-исследовательского проекта:**
	Си1. Применение прогнозной модели позволяет улучшить стратегическое планирование в нефтегазовой отрасли Си2. Применение прогнозной модели на основе больших данных позволяет оптимизировать процессы и принимать более информированные решения Си3.Небольшая цена разработки, по сравнению с конкурентами	Сл1. Сильная зависимости результатов оценки от качества входных данных Сл2. Небольшой размер входной выборки Сл3. Результат моделирования носит вероятностный характер

Продолжение таблицы 6

Возможности: В1. Улучшение точности и надежности В2. Увеличение эффективности операций В3. Появление дополнительного спроса на программный продукт	Из данной комбинации можем сделать вывод, что существуют возможности улучшить качество модели, что приведет к автоматизации и оптимизации бизнес-процессов в нефтедобыче.	Для достижения высокого качества модели необходимы выборки больших объемов, но обычно такие данные требуют серьезной обработки и могут иметь высокую цену сбора.
Угрозы: У1. Уход от использования нефтепродуктов У2. Появление более качественных аналогов модели У3. Появление более дешёвых аналогов	Используемые подходы и алгоритмы могут быть применимы к другим областям при грамотной адаптации.	Необходимость в постоянном улучшение модели в связи с ее вероятностным характером не позволяет поддерживать стабильную ценовую политики.

Самым большим преимуществом данной разработки является еѐ гибкая архитектура, а недостатком – сильная зависимость от предоставляемых пользователем данных для корректной работы и работы облачных сервисов.

5.2 Инициация проекта

В данном разделе приводится информация о заинтересованных сторонах проекта, иерархии целей проекта и критериях достижения целей. Информация о заинтересованных сторонах проекта представлена в таблице 7.

Таблица 7 – Заинтересованные стороны проекта

Заинтересованные стороны	**Ожидания**
Компания-пользователь	Оценка эффективности скважин от входных параметров
Руководитель, Исполнитель	Выполненная выпускная квалификационная работа

В таблице 8 представлена рабочая группа проекта, определена роль каждого участника в данном проекте, а также прописаны функции, выполняемые участниками.

Таблица 8 – Рабочая группа проекта

№	ФИО, основное место работы, должность	Роль в разработке	Функции
1	Губин Е. И., ТПУ ОИТ ИШИТР, доцент	Руководитель	Утверждение основных разделов, выдача заданий к использованию, координирование деятельности исполнителя
2	Еременко Марк Сергеевич, ТПУ ОИТ ИШИТР, магистрант гр. 8ПМ1И	Исполнитель	Выполнение поставленных задач

Цели и результат проекта представлены в таблице 9.

Таблица 9 – Цели и результат проекта

Цели проекта	● Изучить предметную область ● Получить данные ● Провести подготовку исходных данных ● Прогнозное моделирование ● Оценка результатов
Ожидаемые результаты	● Хорошая оценка модели, высокая вероятность внедрения в технологический процесс, существенная экономия финансовых затрат при внедрении данной методики ● Сдана выпускная квалификационная работа
Критерии приёмки	● Полученная модель адекватно предсказывает наиболее значимые входные данные, влияющие на добычу нефти в предлагаемых условиях
Требования к результату проекта	● Выполнены все пункты функционального требования

5.3 Планирование научно-исследовательских работ

5.3.1 Структура работ в рамках научного исследования

В таблице 10 приведен порядок работ, выполняемых в ходе разработки, и исполнитель каждой работы.

Таблица 10 – Перечень работ и исполнителей при разработке модуля

№	Наименование работы	Исполнители работы
1	Выбор научного руководителя работы	Руководитель

Продолжение таблицы 10

2	Составление и утверждение темы работы	Руководитель Исполнитель
3	Составление календарного плана-графика выполнения работы	Руководитель Исполнитель
4	Подбор и изучение литературы по теме работы	Исполнитель
5	Анализ предметной области	Исполнитель
6	Формирование выборок, конструирование признаков, очистка, интеграция и форматирование данных.	Исполнитель
7	Выбор, обучение и оценка качества модели.	Исполнитель
8	Оценка процесса, полученных результатов и определение последующих действий.	Руководитель Исполнитель
9	Согласование выполненной работы с научным руководителем	Руководитель Исполнитель
10	Выполнение других частей работы (финансовый менеджмент, социальная ответственность)	Исполнитель
11	Подведение итогов, оформление работы	Руководитель Исполнитель

5.3.2 Бюджет научно-технического исследования

В процессе формирования бюджета научно-технического исследования используется следующая группировка затрат по статьям:

— материальные затраты НТИ;

— затраты на основное оборудование;

— основная заработная плата исполнителей темы;

— дополнительная заработная плата исполнителей темы;

— отчисления во внебюджетные фонды (страховые отчисления);

— накладные расходы.

5.3.2.1 Расчет материальных затрат на научное исследование

В материальные затраты научного исследования вошли затраты на канцелярские принадлежности, оплата интернет провайдера (таблица 11).

Таблица 11 – Расчет затрат на по статье «Материальные затраты на научное исследование»

Наименование	Марка, размер	Кол-во	Цена за единицу, руб.	Сумма, руб.
Бумага офисная	А4 Снегурочка 80 г/м²	500 листов	0,598	299
Интернет	Ростелеком, стандартный пакет	5 месяцев	409	2 045
Ручка	Ручка “Erich Krause”	1 шт.	98	98
Всего за материалы				2 442
Транспортно-заготовительные расходы (3-5%)				122
Итого по статье $C_м$				2 564

Таким образом, общая сумма материальных затрат составляет 2564,1 рублей.

5.3.2.2 Специальное оборудование для научных работ

Поскольку в ходе работы использовалось программное обеспечение с открытым исходным кодом, затраты на оборудование включают в себя только затраты на оборудование студента.

Во время проведения научного исследования использовался ПК стоимостью 80000 рублей. Расчёт затрат на амортизацию представлен в таблице 12.

Таблица 12 – Расчет затрат на по статье «Спецоборудование для научных работ»

п/п	Наименование оборудования	Кол-во единиц оборудования	Цена единицы оборудования, тыс. руб	Общая стоимость оборудования, тыс. руб
	Ноутбук	1	80 000	80 000
Итого				80 000

Таким образом, общая сумма затрат на специальное оборудование для научных работ составляет 80000 рублей.

5.3.2.3 Основная и дополнительная заработная плата

В настоящую статью включается основная заработная плата и сумма выплат, предусмотренных законодательством о труде, научных и инженерно-технических работников, рабочих макетных мастерских и опытных производств, непосредственно участвующих в выполнении работ по данной теме. Величина расходов по заработной плате определяется исходя из трудоемкости выполняемых работ и действующей системы окладов и тарифных ставок. В состав основной заработной платы включается премия, выплачиваемая ежемесячно из фонда заработной платы в размере 20 – 30 % от тарифа или оклада.

Статья включает основную заработную плату работников, непосредственно занятых выполнением НТИ, (включая премии, доплаты) и дополнительную заработную плату:

$$З_{зп} = З_{осн} + З_{доп}, \quad (3)$$

где $З_{осн}$ – основная заработная плата;

$З_{доп}$ – дополнительная заработная плата (12-20 % от $З_{осн}$).

Основная заработная плата ($З_{осн}$) руководителя (лаборанта, инженера) от предприятия (при наличии руководителя от предприятия) рассчитывается по следующей формуле:

$$З_{осн} = З_{дн} \cdot Т_{р}, \quad (4)$$

где $З_{осн}$ – основная заработная плата одного работника;

$Т_{р}$ – продолжительность работ, выполняемых научно-техническим работником, раб. дн.;

$З_{дн}$ – среднедневная заработная плата работника, руб.

Среднедневная заработная плата рассчитывается по формуле:

$$З_{дн} = \frac{З_{м} \cdot М}{F_{д}}, \quad (5)$$

где $З_{м}$ – месячный должностной оклад работника, руб.;

$F_{д}$ – количество рабочих дней в месяце (среднее количество рабочих дней – 25);

М – количество месяцев работы без отпуска в течение года: при отпуске в 48 раб. дней М=10,4 месяца, 6-ти дневная неделя.

В таблице 13 приведён баланс рабочего времени для участников разработки.

Таблица 13 – Баланс рабочего времени участников разработки

Показатели рабочего времени	**Руководитель**	**Исполнитель**
Календарное число дней	365	365
Количество нерабочих дней - выходные дни - праздничные дни	104 14	104 14
Потери рабочего времени - отпуск - невыходы по болезни	48 –	24 –
Действительный годовой фонд рабочего времени	199	223

Дополнительная заработная плата рассчитывается исходя из 10-15% от основной заработной платы, работников, непосредственно участвующих в выполнение темы:

$$З_{доп} = k_{доп} \cdot З_{осн}, \quad (6)$$

где $k_{доп}$ – коэффициент дополнительной заработной платы;

$З_{доп}$ – дополнительная заработная плата, руб.;

$З_{осн}$ – основная заработная плата, руб.

Данные для расчета: оклад у научного руководителя (доцент) – 33 162 руб., оклад у исполнителя (инженер) – 14 874 руб.

1. Действительный годовой фонд рабочего времени научно-технического персонала (руководитель – 199дн., инженер – 223дн.).
2. Коэффициент дополнительной заработной платы 15%
3. Районный коэффициент – 30%;

Определяем основную заработную плату для исполнителя:

$$З_{дн} = \frac{14\,874 \cdot 10{,}4}{223} = 694\ руб./день$$

$$З_{осн} = 694 \cdot 114 = 79\,079\ руб.$$

Дополнительная заработная плата исполнителя:

$$З_{доп} = 0{,}15 \cdot 79\,079 = 11\,862 руб.$$

Итого затраты на оплату труда:

$$З_{общ} = 79\,079 + 11\,862 = 90\,941\ руб.$$

Общая сумма заработной платы с учетом районного коэффициента:

$$З_{общ} = 90\,941 \cdot 1{,}3 = 118\,223\ руб.$$

Теперь рассчитаем основную заработную плату руководителя:

$$З_{дн} = \frac{33\,162 \cdot 10{,}4}{199} = 1\,733\ руб./день$$

$$З_{осн} = 1\,733 \cdot 8 = 13\,865\ руб.$$

Дополнительная заработная плата руководителя:

$$З_{доп} = 0{,}15 \cdot 13\,865 = 2\,080\ руб.$$

Итого затраты на оплату труда:

$$З_{общ} = 13\,865 + 2\,080 = 15\,945\ руб.$$

Общая сумма заработной платы с учетом районного коэффициента:

$$З_{общ} = 15\,945 \cdot 1{,}3 = 20\,728\ руб.$$

Тогда, общая сумма затрат на заработную плату составит:

$$З_{общ} = 118\,223 + 20\,728 = 138\,951 \text{ руб.}$$

Расчет затраты на основную заработную плату сведен в таблицу 14.

Таблица 14 – Затраты на основную заработную плату

Исполнитель	Оклад (руб.)	Среднедневная заработная плата (руб./день)	Основная заработная плата (руб.)	Дополнительная заработная плата (руб.)	Заработная плата с учетом районного коэффициента (руб.)
1.Руководитель	33 162	1 733	13 865	2 080	20 728
2. Исполнитель	14 874	694	79 079	11 862	118 223
Итого:					138 951

5.3.2.4 Отчисления на социальные нужды

В данной статье расходов отражаются обязательные отчисления по установленным законодательством Российской Федерации нормам органам государственного социального страхования (ФСС), пенсионного фонда (ПФ) и медицинского страхования (ФФОМС) от затрат на оплату труда работников.

Величина отчислений во внебюджетные фонды определяется исходя из следующей формулы:

$$C_{\text{внеб}} = k_{\text{внеб}} \cdot З_{\text{общ}}, \quad (7)$$

$$C_{\text{внеб}} = 0{,}3 \cdot 138\,951 = 41\,685\ руб.$$

где $k_{\text{внеб}}$ = 30 % – коэффициент отчислений на уплату во внебюджетные фонды (пенсионный фонд, фонд обязательного медицинского страхования и пр.);

5.3.2.5 Накладные расходы

Накладные расходы учитывают прочие затраты организации, не попавшие в предыдущие статьи расходов: печать и ксерокопирование материалов исследования, оплата услуг связи, электроэнергии, почтовые и телеграфные расходы, размножение материалов и т.д.

Накладные расходы составляют 80-100 % от суммы основной и дополнительной заработной платы, работников, непосредственно участвующих в выполнение темы.

Расчет накладных расходов ведется по следующей формуле:

$$C_{\text{накл}} = k_{\text{накл}} \cdot (З_{\text{осн}} + З_{\text{доп}}), \quad (8)$$

где $k_{\text{накл}}$ – коэффициент накладных расходов;

$$C_{\text{накл}} = 0{,}8 \cdot 138\,951 = 111\,161\ руб.$$

5.3.2.6 Формирование бюджета затрат научно-исследовательского проекта

Рассчитанная величина затрат научно-исследовательской работы (темы) является основой для формирования бюджета затрат проекта, который при

формировании договора с заказчиком защищается научной организацией в качестве нижнего предела затрат на разработку научно-технической продукции.

Определение бюджета затрат на научно-исследовательский проект по каждому варианту исполнения приведен в таблице 15.

Таблица 15 – Расчет бюджета затрат НТИ

Наименование статьи	Сумма, руб.
1. Материальные затраты НТИ	2 564
2. Затраты на специальное оборудование для научных (экспериментальных) работ	80 000
3. Затраты по основной заработной плате исполнителей темы	138 951
4. Отчисления во внебюджетные фонды	41 685
5. Накладные расходы	111 161
Общий бюджет затрат НТИ	374 361

По итогу расчётов имеем общий бюджет проекта 374 361 руб., с учетом затрат по заработной плате исполнителей 138 951руб.

5.4 Определение ресурсной (ресурсосберегающей), финансовой, бюджетной, социальной и экономической эффективности исследования

Определение эффективности происходит на основе расчета интегрального показателя эффективности научного исследования. Его нахождение связано с определением двух средневзвешенных величин: финансовой эффективности и ресурсоэффективности.

Интегральный показатель финансовой эффективности научного исследования получают в ходе оценки бюджета затрат трех (или более) вариантов исполнения научного исследования. Для этого наибольший интегральный показатель реализации технической задачи принимается за базу

расчета (как знаменатель), с которым соотносится финансовые значения по всем вариантам исполнения.

Интегральный финансовый показатель разработки определяется как:

$$I_{\text{финр}}^{\text{ипс}.i} = \frac{\Phi_{\text{р}i}}{\Phi_{max}}, \quad (9)$$

$$I_{\text{финр}}^{\text{ипс}.i} = \frac{374\,361}{374\,361} = 1$$

где $I_{\text{финр}}^{\text{ипс}.i}$ – интегральный финансовый показатель разработки;

$\Phi_{\text{р}i}$ – стоимость *i*-го варианта исполнения;

Φ_{max} – максимальная стоимость исполнения научно-исследовательского проекта (в т.ч. аналоги).

Полученная величина интегрального финансового показателя разработки отражает соответствующее численное увеличение бюджета затрат разработки в разах (значение больше единицы), либо соответствующее численное удешевление стоимости разработки в разах (значение меньше единицы, но больше нуля).

Интегральный показатель ресурсоэффективности вариантов исполнения объекта исследования можно определить следующим образом:

$$I_{pi} = \sum a_i \cdot b_i, \quad (10)$$

где I_{pi} – интегральный показатель ресурсоэффективности для *i*-го варианта исполнения разработки;

a_i – весовой коэффициент *i*-го варианта исполнения разработки;

b_i^a, b_i^p – бальная оценка *i*-го варианта исполнения разработки, устанавливается экспертным путем по выбранной шкале оценивания;

n – число параметров сравнения.

Расчет интегрального показателя ресурсоэффективности рекомендуется проводить в форме таблицы (таблица 16).

Таблица 16 – Сравнительная оценка характеристик вариантов исполнения проекта

Объект исследования Критерии	Весовой коэффициент параметра	Исп.1
1. Производительность	0,3	5
2. Энергосбережение	0,4	5
3. Надежность	0,2	4
4. Материалоемкость	0,1	4
ИТОГО	1	

$$I_{p-\text{исп1}} = 5 \cdot 0{,}3 + 5 \cdot 0{,}4 + 4 \cdot 0{,}2 + 4 \cdot 0{,}1 = 4{,}7 \qquad (11)$$

Исходя из полученных результатов, можно сделать вывод, что научно-техническое исследование будет ресурсоэффективно при первом варианте исполнения проекта, т.е. при высокой производительности и энергоэффективности проекта.

Интегральный показатель эффективности вариантов исполнения разработки ($I_{\text{исп}\,i}$) определяется на основании интегрального показателя ресурсоэффективности и интегрального финансового показателя по формуле:

$$I_{\text{исп.1}} = \frac{I_{p-\text{исп1}}}{I_{\text{фин}p}^{\text{исп.1}}}, I_{\text{исп.2}} = \frac{I_{p-\text{исп2}}}{I_{\text{фин}p}^{\text{исп.2}}} \text{ и т. д.} \qquad (12)$$

Сравнение интегрального показателя эффективности вариантов исполнения разработки позволит определить сравнительную эффективность проекта (таблица 17) и выбрать наиболее целесообразный вариант из предложенных. Сравнительная эффективность проекта ($Э_{\text{ср}}$):

$$Э_{\text{ср}} = \frac{I_{\text{исп1}}}{I_{\text{исп2}}} \qquad (13)$$

Таблица 17 – Сравнительная эффективность разработки

№ п/п	Показатели	Исп.1
1	Интегральный финансовый показатель разработки	1
2	Интегральный показатель ресурсоэффективности разработки	4,7
2	Интегральный показатель эффективности	0,21

Сравнение значений интегральных показателей эффективности позволяет понять и выбрать более эффективный вариант решения поставленной в бакалаврской работе технической задачи с позиции финансовой и ресурсной эффективности.

5.5 Вывод по разделу

В ходе данной работе были рассмотрены потенциальные потребители результатов исследования, также для анализа конкурентных технических решений. С позиции ресурсоэффективности и ресурсосбережения была составлена оценочная карта сравнения конкурентных технических решений, по результату которой разрабатываемая система имеет лучшие качества.

Далее был сформирован SWOT-анализ, в котором балы описаны сильные и слабые стороны проекта, в выявлении возможностей и угроз для реализации проекта, для выявления соответствия и несоответствия была составлена интерактивная матрица проекта.

В рамках процессов инициации определены внутренние и внешние заинтересованные стороны проекта с их ожиданиями от проекта, цели и результат проекта.

План проекта представлен на диаграмме Ганта, из которого какие виды работ осуществлялись руководителем и исполнителем проекта в течение какого количества дней.

В бюджет инженерно-технического проекта занесены материальные затраты в размере 2564 рублей. Также добавлена стоимость оборудования, которая составила 80000 рублей. Была рассчитана основная и дополнительная заработная плата исполнителей проекта, сумма которых составила 138 951 рублей. Общий бюджет проекта получился равен 374 361 рублям.

Проведена оценка результатов ресурсоэффективности, которая составила 4,7 из 5, что говорит о хорошей эффективности реализации технического проекта.

6 Социальная ответственность

ЗАДАНИЕ ДЛЯ РАЗДЕЛА «СОЦИАЛЬНАЯ ОТВЕТСТВЕННОСТЬ»

Обучающемуся:

Группа	ФИО		
8ПМ1И	Еременко Марк Сергеевич		
Школа	**ИШИТР**	**Отделение (НОЦ)**	**Отделение информационных технологий**
Уровень образования	магистратура	**Направление/ специальность**	09.04.04 «Программная инженерия»

Тема ВКР:

Исходные данные к разделу «Социальная ответственность»:	
Введение – Характеристика объекта исследования (вещество, материал, прибор, алгоритм, методика) и области его применения. – Описание рабочей зоны (рабочего места) при разработке проектного решения/при эксплуатации	*Объект исследования:* прогнозная модель для оценки эффективности скважин. *Область применения:* разработка нефтяных месторождений. *Рабочая зона:* офис. *Размеры помещения: 5x3x4 м.* *Количество и наименование оборудования рабочей зоны:* рабочий стол, рабочее кресло и персональный компьютер. *Рабочие процессы, связанные с объектом исследования, осуществляющиеся в рабочей зоне:* анализ литературы, сбор исходных данных, разработка алгоритма.
Перечень вопросов, подлежащих исследованию, проектированию и разработке:	
1. Правовые и организационные вопросы обеспечения безопасности при разработке проектного решения: – специальные (характерные при эксплуатации объекта исследования, проектируемой рабочей зоны) правовые нормы трудового законодательства; – организационные мероприятия при компоновке рабочей зоны.	СП 2.4.3648-20. «Санитарно-эпидемиологические требования к организациям воспитания и обучения, отдыха и оздоровления детей и молодежи»; ГОСТ Р 50923-96. Дисплеи. Рабочее место оператора; ГОСТ 12.2.032-78. Рабочее место при выполнении работ сидя.

2. Производственная безопасность при разработке проектного решения: – Анализ выявленных вредных и опасных факторов; – Расчет уровня опасного или вредного производственного фактора	**Опасные факторы:** 1. Повышенное значение напряжения в электрической цепи, замыкание которой может произойти через тело человека. **Вредные факторы:** 1. Отсутствие или недостаток естественного света; 2. Недостаточная освещенность рабочей зоны; 3. Перенапряжение анализаторов; 4. Превышение уровня шума на рабочем месте; 5. Повышенный уровень электромагнитных излучений; 6. Статические перегрузки, связанные с рабочей позой. **Требуемые средства коллективной и индивидуальной защиты от выявленных факторов:** очки со стеклами, отражающими синее излучение от ЖК-мониторов. **Расчет:** расчет системы искусственного освещения.
3. Экологическая безопасность при разработке проектного решения и эксплуатации: – Разработка нефтяных месторождений.	**Воздействие на литосферу:** перекрытие почвенного профиля застройками нефтяных скважин. Разлитие буровых растворов, химических агентов и нефти, нарушение естественного залегания пород. **Воздействие на гидросферу:** загрязнение производственных сточных вод, используемых в технологических процессах в нефтеперерабатывающей промышленности, загрязнение водотоков, подземных грунтовых вод химическими реагентами, отходами и нефтью. **Воздействие на атмосферу:** сброс газа на факельное устройство, выбросы, выхлопные газы **Воздействие на селитебную зону:** загрязнение химическими веществами
4. Безопасность в чрезвычайных ситуациях при разработке проектного решения:	**Возможные ЧС:** Природная: сильные морозы зимой (аварии на электро-, тепло-коммуникациях, водоканале, транспорте); Техногенная: несанкционированное проникновение посторонних на рабочее место (возможны проявления вандализма, диверсии, промышленного шпионажа). **Наиболее типичная ЧС:** возникновение пожара.
Дата выдачи задания для раздела по линейному графику	

Задание выдал консультант:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ООД ШБИП	Антоневич Ольга Алексеевна	к.б.н.		

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
8ПМ1И	Еременко Марк Сергеевич		

6 Социальная ответственность

Целью данной выпускной квалификационной работы является разработка методики анализа данных и построение прогнозной модели для выявления перспективных месторождений. Основные проблемы, решения которых требует социальная ответственность состоят в том, чтобы обеспечить нормальные условия деятельности людей, и защитить человека и окружающую его среду от воздействия вредных и опасных факторов, превышающих нормативно-допустимые уровни.

Поддержка хороших условий труда и отдыха человека создаёт условия для высокой работоспособности и продуктивности. Обеспечение безопасности труда и отдыха способствует сохранению жизни и здоровья людей за счет снижения травматизма и заболеваемости. Поэтому объектом изучения социальной ответственности является комплекс отрицательно воздействующих факторов и процессов на человека в рабочей среде.

Целью данного раздела является выявить и проанализировать рабочее место на наличие опасных и вредных факторов, а также принять меры по ограничению их воздействия на человека. В качестве организационного вопроса обеспечения безопасности присутствует необходимость соблюдения норм, инструкций и прочих документов, утвержденных в порядке законом.

6.1 Правовые и организационные вопросы обеспечения безопасности

6.1.1 Правовые нормы трудового законодательства

Режим рабочего времени сотрудника и установления его перерывов во время работы за компьютером нормативно не урегулирован, согласно СП 2.4.3648-20 «Санитарно-эпидемиологические требования к организациям воспитания и обучения, отдыха и оздоровления детей и молодежи» [19]. Работодатель может самостоятельно установить порядок предоставления перерывов в работе за компьютером для отдыха в правилах внутреннего трудового распорядка. Указанные перерывы включаются в рабочее время. То

есть они не продлевают продолжительность рабочего дня сотрудника. Согласно трудовому кодексу Российской Федерации от 30.12.2001 г. N 197–ФЗ (ТК РФ) во время этих перерывов работник не должен выполнять другую работу [20].

6.1.2 Эргономические требования к правильному расположению и компоновке рабочей зоны

Работа построения прогнозных моделей для выявления перспективных нефтяных месторождений подразумевает, что сотрудник осуществляет данное действие, сидя за персональным компьютером.

Конструкция рабочего места и взаимное расположение всех его элементов (стул, стол, средства отображения информации и т.д.) должны соответствовать общим эргономическим требованиям, приведённым в таблице 18 [21].

Таблица 18 – Нормы оборудования рабочих мест

Ширина рабочего стола	не менее 600 мм
Глубина рабочего стола	не менее 1200 мм
Высота рабочего стола	От 680 до 800 мм (если высота стола не регулируется – 725 мм)
Угол наклона спинки	в пределах 0°±30° от вертикального положения
Расстояние спинки от переднего края сидения	от 260 до 400 мм
Высота поверхности сидения	от 400 до 550 мм
Сидение	Ширина и глубина не менее 400 мм
Подставка для ног	Ширина – от 300 мм, глубина – от 400 мм, с углом наклона до 20 градусов
Расстояние клавиатуры от края стола	от 100 до 300 мм

Согласно ГОСТ Р 50923-96. «Дисплеи. Рабочее место оператора» [21] и ГОСТ 12.2.032-78. «Рабочее место при выполнении работ сидя» [22], монитор на рабочем месте оператора должен располагаться так, чтобы изображение в любой его части было различимо без необходимости поднять или опустить

голову. Дисплей на рабочем месте должен быть установлен ниже уровня глаз оператора. Угол наблюдения экрана оператором относительно горизонтальной линии взгляда не должен превышать 60°.

Клавиатура на рабочем месте оператора должна располагаться так, чтобы обеспечивалась оптимальная видимость экрана. Также клавиатура должна иметь возможность свободного перемещения.

6.2 Производственная безопасность

В данном разделе проанализированы вредные и опасные факторы, которые могут возникать при разработке проектного решения.

Перечень опасных и вредных факторов, характерных для объекта исследования представлен в таблице 19 согласно ГОСТ 12.0.003-2015 «Система стандартов безопасности труда (ССБТ). Опасные и вредные производственные факторы. Классификация.» [23].

Таблица 19 – Возможные вредные и опасные факторы на рабочем месте

Факторы (ГОСТ 12.0.003-2015)	Нормативные документы
Вредные факторы	
1. Отсутствие или недостаток естественного света	Трудовой кодекс Российской Федерации от 30.12. 2001 г. № 197–ФЗ (ред. от 01.04.2019 г.) [19]. СП 52.13330.2016. Естественное и искусственное освещение [24]. СанПиН 1.2.3685-21. Гигиенические нормативы и требования к обеспечению безопасности и (или) безвредности для человека факторов среды обитания [25]. ГОСТ Р 50923-96. Дисплеи. Рабочее место оператора [21].
2. Недостаточная освещенность рабочей зоны	
3. Перенапряжение анализаторов	

Продолжение таблицы 19

4. Превышение уровня шума на рабочем месте	ГОСТ Р 50948-2001. Средства отображения информации индивидуального пользования. Общие эргономические требования и требования безопасности [26]. ГОСТ Р ИСО 9241-5-2009. Эргономические требования к проведению офисных работ с использованием видеодисплейных терминалов [27]. ГОСТ 12.1.003-2014 ССБТ. Шум. Общие требования безопасности [28].
5. Повышенный уровень электромагнитных излучений	
6. Статические перегрузки, связанные с рабочей позой	
Опасные факторы	
1. Повышенное значение напряжения в электрической цепи, замыкание которой может произойти через тело человека	ГОСТ 12.1.038-82 ССБТ. Предельно допустимые значения напряжений прикосновения и токов [29].

6.2.1 Отсутствие или недостаток естественного света и недостаточная освещенность рабочей зоны

Помещения должны иметь как естественное, так и искусственное освещение. Согласно СП 52.13330.2016. «Естественное и искусственное освещение» естественное освещение осуществляется через светопроемы, обеспечивающие необходимый коэффициент естественной освещенности (КЕО) не ниже 1,2 % [24].

1) Источник возникновения фактора – вредное воздействие параметров освещения проявляется в отсутствии или недостатке естественного света, а также недостаточной освещенности рабочей зоны;
2) Воздействие фактора на организм человека – недостаточное освещение влияет на функционирование зрительного аппарата, то есть определяет зрительную работоспособность, на психику человека, его эмоциональное состояние, вызывает усталость

центральной нервной системы, возникающей в результате прилагаемых усилий для опознания четких или сомнительных сигналов;

3) Допустимые нормы: согласно СанПиН 1.2.3685-21 «Гигиенические нормативы и требования к обеспечению безопасности и (или) безвредности для человека факторов среды обитания» освещенность рабочего места в кабинетах, рабочих комнатах и офисах при работе за ЭВМ в горизонтальной плоскости от общего искусственного освещения должна быть 300 [24];
4) Предлагаемые средства защиты – к средствам нормализации освещенности рабочих мест относятся: источники света, осветительные приборы, световые проемы, светозащитные устройства, светофильтры, защитные очки.

Ниже приведены расчеты для создания освещенности E = 300 ЛК для стандартного офисного помещения.

Размеры помещения: A = 4,8 м, ширина B = 4,3 м, высота H = 2,8 м. Высота рабочей поверхности h*pn* = 0,8 м.

Коэффициент отражения стен Rc = 30%, потолка Rn = 50 %. Коэффициент запаса k = 1,5, коэффициент неравномерности Z = 1,1. Рассчитываем систему общего люминесцентного освещения.

Для заданной высоты помещения подойдут двухламповые светильники ШОД с λ = 1,2.

Приняв h*c* = 0,5 м, определяем расчетную высоту:

$$h = H - hc - hpn = 2{,}8 - 0{,}5 - 0{,}8 = 1{,}5 \text{ м}; \quad (14)$$

Расстояние между светильниками:

$$L = \lambda \cdot h = 1{,}5 \cdot 1{,}2 = 1{,}8 \text{ м}; \quad (15)$$

Расстояние от крайнего ряда светильников до стены:

$$L / 3 = 0{,}6 \text{ м}. \quad (16)$$

Размещаем светильники в три ряда. В каждом ряду можно установить 2 светильника типа ШОД мощностью 40 Вт (с длиной 1,228 м), при этом

разрывы между светильниками в ряду составят 1,144 м. Изображаем в масштабе план помещения и размещения на нем светильников (рисунок 14). Учитывая, что в каждом светильнике установлено две лампы, общее число ламп в помещении должно быть N = 12.

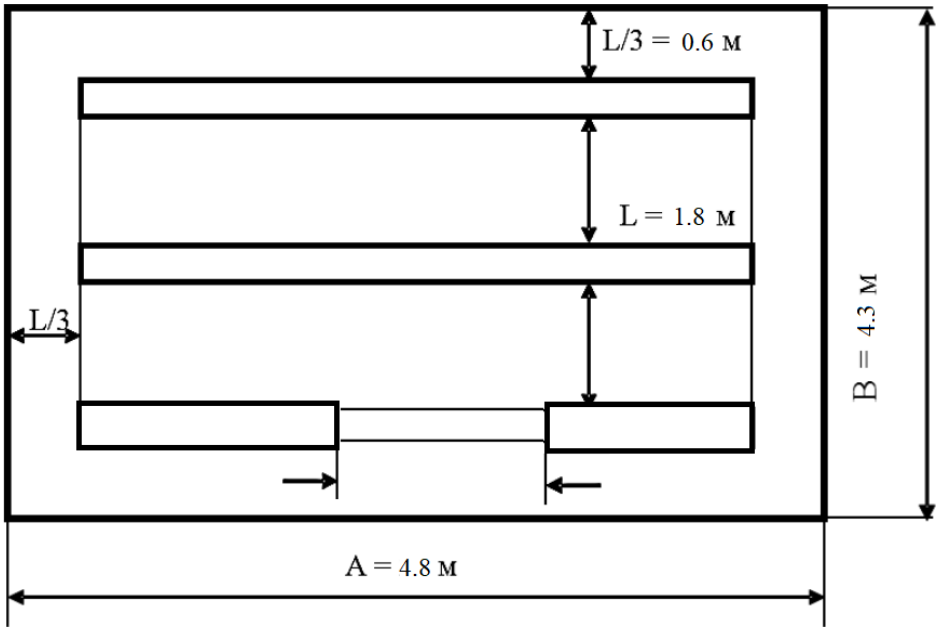

Рисунок 14 – План помещения и размещения светильников с люминесцентными лампами

6.2.2 Перенапряжение анализаторов

Основная характеристика анализаторов – высокая чувствительность, хотя не всякий раздражитель, действующий на анализатор, вызывает ощущение. Чтобы ощущение проявилось, необходима определенная интенсивность раздражителя. Всякое воздействие, превышающее предел интенсивности, вызывает боль и нарушение деятельности анализаторов.

1) Источник возникновения фактора – поступающая информация с монитора компьютера;
2) Воздействие фактора на организм человека – интенсивное или длительное воздействие перенапряжение анализаторов может привести к функциональному чрезмерному напряжению, стать причиной профессиональных заболеваний;
3) Допустимые нормы – с 2021 года вопрос установления перерывов во время работы за компьютерами нормативно не урегулирован.

Работодатель может самостоятельно установить порядок предоставления перерывов в работе за компьютером для отдыха в правилах внутреннего трудового распорядка. Указанные перерывы включаются в рабочее время. То есть они не продлевают продолжительность рабочего дня сотрудника. Во время этих перерывов работник не должен выполнять другую работу. Перерыв предоставляется ему для отдыха. Также перерывы в работе для отдыха от компьютера нужно предоставлять отдельно от перерыва на обед согласно трудовому кодексу Российской Федерации [20].

4) Предлагаемыми средствами защиты для минимизации воздействия фактора являются регулярные перерывы для сотрудников, работающих с данным проектным решением.

6.2.3 Превышение уровня шума на рабочем месте

Одним из наиболее распространенных в производстве вредных факторов является шум. Он создается вентиляционным и рабочим оборудованием, преобразователями напряжения, рабочими лампами дневного света, а также проникает снаружи. Шум вызывает головную боль, усталость, бессонницу или сонливость, ослабляет внимание, память ухудшается, реакция уменьшается.

Основным источником шума в комнате являются компьютерные охлаждающие вентиляторы и. Уровень шума варьируется от 35 до 42 дБА [28].

При значениях выше допустимого уровня необходимо предусмотреть средства индивидуальной защиты (СИЗ) и средства коллективной защиты (СКЗ) от шума.

Средства коллективной защиты:

1. Устранение причин шума или существенное его ослабление в источнике образования;
2. Изоляция источников шума от окружающей среды (применение глушителей, экранов, звукопоглощающих строительных

материалов, например, любой пористый материал – шамотный кирпич, микропористая резина, поролон и др.);

3. Применение средств, снижающих шум и вибрацию на пути их распространения.

6.2.4 Повышенный уровень электромагнитных излучений

1. Источник возникновения фактора – дисплеи (мониторы). Они представляют собой источники наиболее вредных излучений, неблагоприятно влияющих на здоровье человека;
2. Воздействие фактора на организм человека – при длительном воздействии данного фактора возникают жалобы на слабость, раздражительность, быструю утомляемость и ослабление памяти;
3. Допустимые нормы: уровни электромагнитного поля приведены в таблице 20 согласно ГОСТ Р 50948-2001 «Средства отображения информации индивидуального пользования. Общие эргономические требования и требования безопасности.» [26].
4. Предлагаемые средства защиты – рациональное размещение оборудования; использование средств, ограничивающих поступление электромагнитной энергии на рабочие места персонала (экраны-фильтры и защитные очки).

Таблица 20 – Допустимые уровни электромагнитного поля

Наименование параметров		**ВДУ ЭМП**
Напряженность электрического поля	В диапазоне частот 5 Гц – 2 кГц	25 В/м
	В диапазоне частот 2 кГц – 400 кГц	2,5 В/м
Плотность магнитного потока	В диапазоне частот 5 Гц – 2 кГц	250 нТл
	В диапазоне частот 2 кГц – 400 кГц	25 нТл
Электростатический потенциал экрана видеомонитора		500 В

6.2.5 Статические перегрузки, связанные с рабочей позой

1) Источник возникновения фактора – рабочее место;
2) Воздействие фактора на организм человека – неправильная рабочая поза может привести к хроническому спазму (повышенной напряженности) мышц руки, невралгии, плекситу, обострению шейного и грудного радикулита и ряду других неврологических заболеваний;
3) Допустимые нормы согласно ГОСТ Р ИСО 9241-5-2009 «Эргономические требования к проведению офисных работ с использованием видео дисплейных терминалов»:

– бедра расположены приблизительно в горизонтальной позиции, а ноги от колена до ступни - в вертикальной позиции; высота сиденья должна равняться длине голеней пользователя до подколенной области или быть немного меньше;

– плечо расположено вертикально, предплечье – горизонтально;

– работа не требует сгибаний или разгибаний запястий;

– позвоночник расположен вертикально;

– ступня составляет угол в 90° по отношению к подколенной части ноги;

– скручивание верней части туловища отсутствует;

– линия зрения заключена между горизонталью и 60° ниже горизонтали.

4) Предлагаемые средства защиты – правильная организация рабочего места.

6.2.6 Повышенное значение напряжения в электрической цепи, замыкание которой может произойти через тело человека

В работе с электрооборудованием может возникнуть повышенное значение напряжения в электрической цепи, замыкание которой может произойти через тело человека.

1) Источник возникновения фактора – электрическое оборудование (ПК, промышленный контроллер);

2) Воздействие фактора на организм человека – электрический ток, проходя через организм, оказывает термическое, электролитическое и биологическое действие. Термическое действие выражается в ожогах отдельных участков тела, нагреве кровеносных сосудов, нервов и других тканей. Электролитическое действие выражается в разложении крови и других органических жидкостей, что вызывает значительные нарушения их физико-химических составов. Биологическое действие выражается в раздражении и возбуждении живых тканей организма, а также в нарушении внутренних биоэлектрических процессов, протекающих в нормально действующем организме и теснейшим образом связанных с его жизненными функциями;

3) Допустимые нормы: согласно ГОСТ 12.1.038-82 ССБТ “Предельно допустимые значения напряжений прикосновений и токов” номинальное напряжение не превышает 50 В переменного тока (действующее значение) или 120 В постоянного (выпрямленного) тока. Напряжения прикосновения и токи, протекающие через тело человека при нормальном (неаварийном) режиме электроустановки, не должны превышать значений, указанных в таблице 21.

Таблица 21 – Допустимые значения напряжений прикосновений и токов

Род тока	U, В	I, мА
	Не более	
Переменный, 50 Гц	2,0	0,3
Переменный, 400 Гц	3,0	0,4
Постоянный	8,0	1,0

4) Предлагаемые средства – для защиты от поражения электрическим током все токоведущие части должны быть защищены от случайных прикосновений кожухами, корпус устройства должен быть заземлен. Заземление выполняется изолированным медным проводом сечением 1.5 мм, который присоединяется к общей шине заземления с общим сечением 48 мм. Общая шина присоединяется к заземлению, сопротивление которого не должно превышать 4 Ом.

Для снижения влияния выявленных опасных и вредных факторов на работающих разработаны следующие мероприятия:

— Организация регулярных перерывов для сотрудников, работающих с данной библиотекой.
— Нормализация освещенности рабочих мест, путем установки дополнительных источников света и осветительных приборов.
— Все электроустановки должны быть снабжены средствами защиты, а также средствами оказания первой помощи в соответствии с действующими правилами применения и испытания средств защиты, используемых в электроустановках.

Применяемое в проектной разработке электрооборудование, электротехнические изделия и материалы должны соответствовать требованиям государственных стандартов или технических условий, утвержденных в установленном порядке согласно ГОСТ Р 12.1.019-2009 «Электробезопасность. Общие требования и номенклатура видов защиты» помещение, в котором используется построение прогнозных моделей, относится к классу помещений без повышенной опасности, в которых отсутствуют условия, создающие повышенную или особую опасность. В данных помещениях с установками напряжением до 1 кВ допускается применение неизолированных и изолированных токоведущих частей без защиты от прикосновения, если по местным условиям такая защита не является необходимой для каких-либо иных целей (например, для защиты от механических воздействий). При этом доступные прикосновению части должны располагаться так, чтобы нормальное обслуживание не было сопряжено с опасностью прикосновения к ним.

Таким образом, рабочее место, в которой проводятся данные проектные работы, соответствует «Правилам устройства электроустановок» и другим нормативам и не требует мероприятий по защите исследователя от действия опасных и вредных факторов.

6.3 Экологическая безопасность

6.3.1. Анализ влияния эксплуатации разработки нефтяных месторождений на окружающую среду

Основными типами воздействий на окружающую среду являются:

— Загрязнение нефтью или химическими реагентами окружающей среды из-за несовершенства технологий или аварийных разливов;

— Загрязнение атмосферы из-за испарений нефтепродуктов;

— Загрязнение отходами промышленного и бытового характера природной среды.

В результате происходит:

— Сокращение ареалов распространения флоры из-за разливов;

— Сокращение рыбных запасов из-за загрязнения поверхностных вод;

— Вырубка лесов из-за обустройства вахтового поселка.

Мерами по охране окружающей среды являются минимизация выброса газа и разлива нефти, а также оптимизация процессов сжигания газов.

6.3.1.1 Воздействие на литосферу

Загрязнение почв нефтью или химическими реагентами приводит к экологическому ущербу, т.е. снижается продуктивность лесов и ухудшается санитарное состояние окружающей среды. Поэтому следует проводить рекультивацию земель.

Рекультивацию загрязненных земель по трассам трубопроводов выполняется следующих образом [30]:

1 этап – происходит сбор пролитой нефти, срез почвенного слоя толщиной 0,2 – 0,4 м и перемещения его во временные отвалы до начала строительных работ;

2 этап – производят поверхностное внесение минеральных удобрений и посев многолетних трав.

Предотвращение аварийных разливов нефти и химических реагентов достигается:

— Контролем за давлением в пласте и оборудовании;

— Аварийным отключением генерирующих агрегатов;

— Контролем за герметичностью оборудования.

6.3.1.2 Воздействие на гидросферу

Разлив нефти, химических реагентов, применяемых при работе генерирующего оборудования, или утилизация остатков реагентов негативно влияют на состав поверхностных вод. При разливе нефти на воде образуется пленка, которая препятствует воздушному обмену.

Пути загрязнения природных вод:

— – При аварии на оборудовании или ее негерметичности могут возникнуть перетоки по затрубному пространству нефти или химических реагентов с последующим попаданием в природные воды;

— – Из-за отсутствия гидроизоляции производственных площадок может произойти загрязнение грунтовых вод. Таким образом, следует не допускать разлива нефти и химических реагентов, чтобы не допустить загрязнения поверхностных и подземных вод.

6.3.1.3 Воздействие на атмосферу

На территории месторождения основными загрязняющими атмосферу веществами являются диоксид азота, окиси углерода, углеводороды, образующиеся в результате сгорания газа на факельных установках, нефти в котельной, углеводороды нефти и попутного газа, обусловленные потерями за счет испарения в системах сбора, хранения и транспорта нефти, а также углеводороды, выделяющиеся в результате залповых выбросов. Залповые выбросы из технологических аппаратов происходят при проверке работоспособности предохранительных клапанов, а также при аварийных ситуациях. Залповые выбросы от технологических аппаратов по существующей технологии направляются на факел сжигания газа [31].

Снижение концентрации загрязняющих веществ в приземном слое атмосферы обеспечивается безаварийной работой технологического оборудования в режиме нормальной эксплуатации промысла.

В целях предупреждения загрязнения атмосферного воздуха, проектными решениями предусматривается ряд мероприятий по сокращению выбросов вредных веществ в атмосферу:

- полная герметизация системы сбора и транспорта нефти и газа;
- стопроцентный контроль швов сварных соединений трубопроводов;
- защита оборудования от коррозии;
- частичная, а в перспективе полная утилизация попутного газа;
- оснащение предохранительными клапанами всей аппаратуры, в которой может возникнуть давление, превышающее расчетное;
- сброс нефти и газа с предохранительных клапанов в аварийные емкости или на факел;
- испытание оборудования на прочность и герметичность после монтажа;
- применение современного блочно-комплексного оборудования заводского изготовления

6.3.1.4 Воздействие селитебную зону

При эксплуатации нефтяных и газовых скважин часты случаи разгерметизации генерирующего оборудования и выбросы различных веществ. Углеводородный газ является вредным для человека, поэтому нельзя допускать загазованности среды.

Метан – токсичен, при недостатке кислорода в воздухе вызывает удушье. Первые признаки отравления – недомогание и головокружение. Присутствие метана в воздухе может привести к пожару и взрыву.

Предельно допустимая концентрация содержания метана в воздухе рабочей зоны – 7000 мг/м3 [32].

Класс опасности – 4 [33].

Метанол – бесцветная прозрачная жидкость по запаху и вкусу напоминает винный (этиловый) спирт - сильный яд, действующий преимущественно на нервную и сосудистую систему. В организм человека

может попасть через дыхательные пути и даже через неповрежденную кожу. Особенно опасен прием метанола внутрь: 5–10 г. метанола может вызвать тяжелое отравление, 30 г. является смертельной дозой.

Симптомы отравления: головная боль, головокружение, тошнота, рвота, боль в желудке, общая слабость, при попадании на слизистую оболочку вызывает раздражение слизистых оболочек.

Метанол при испарении взрывоопасен.

Величина ПДК – 5 (мг/м3) [32].

Класс опасности – 3 [33].

Предельно допустимые концентрации вещества согласно ГОСТ 12.1.005-88: азота диоксид – 2 мг/м3, бензол – 10 мг/м3, углерода оксид – 20 мг/м3., паров нефти – 10 мг/м3

Коллективные средства защиты – устройства, препятствующие появлению человека в опасной зоне. Индивидуальной защиты: очки, защитные маски, противогазы [31].

При расположении генерирующего оборудования вблизи населённых пунктов, необходимо устанавливать датчики анализа среды в разных точках населённого пункта, чтобы не допустить отравления людей.

6.4 Безопасность в чрезвычайных ситуациях

6.4.1 Анализ вероятных ЧС, которые может инициировать объект исследований

Объект исследований может инициировать возникновение такой чрезвычайной ситуации, как пожар. Причинами пожара могут быть неисправность источника питания или компьютера.

6.4.2 Анализ вероятных ЧС, которые могут возникнуть на рабочем месте при проведении проектных работ

При проведении проектных работ также может возникнуть пожар. Причинами пожара могут быть: игнорирование основных правил пожарной

безопасности, неисправность электрической проводки, возгорание устройств искусственного освещения, возгорание устройств вычислительной аппаратуры вследствие нарушения изоляции или неисправности самой аппаратуры.

6.4.3 Обоснование мероприятий по предотвращению ЧС и разработка порядка действия в случае возникновения ЧС

Согласно НПБ 105-03 “Определение категорий помещений, зданий и наружных установок по взрывопожарной и пожарной опасности” помещение, в котором разрабатывалась система, относится к категории В3 по пожароопасности, содержит вещества и материалы, способные гореть при взаимодействии с водой, кислородом воздуха или друг с другом [34].

Помещение содержит ЭВМ, поэтому согласно СП 9.13130.2009 “Техника пожарная. ОГНЕТУШИТЕЛИ. Требования к эксплуатации” для ликвидации пожаров, вызванных возгоранием электрооборудования, применяются углекислотные огнетушители [35].

Для защиты от пожаров необходимо иметь в наличии такое пожарное оборудование как пожарные шкафы, пожарные щиты и огнетушители. Сотрудники должны уметь пользоваться таким оборудованием.

Сотрудники должны знать план эвакуации из помещения, расположение выходов из здания. Также необходимо проводить плановые эвакуации из здания, для того чтобы подготовить сотрудников к действиям в чрезвычайной ситуации.

Чтобы предотвратить пожар в производственном помещении, необходимо:

— работа должна проводиться только при исправном электрооборудовании;

— электросеть не должна перегружаться одновременно несколькими мощными потребителями электроэнергии;

— Уходящий из помещения последним должен проверить выключены ли нагревательные приборы, электроприборы, оборудование и т.д.

При возникновении пожара тушить его самостоятельно целесообразно только на его ранней стадии при обнаружении загорания согласно постановлению Правительства РФ от 16 сентября 2020 г. N 1479 "Об утверждении Правил противопожарного режима в Российской Федерации".

При обнаружении пожара или признаков горения (задымления, запаха гари, повышения температуры) в производственном помещении или на территории предприятия работник обязан немедленно сообщить об этом в пожарную охрану. Пожарной охране сообщается адрес объекта и место возникновения пожара. Сообщить пожарной охране необходимо даже в том случае, если загорание ликвидировано собственными силами. Огонь может остаться незамеченным в скрытых местах (в пустотах деревянных перекрытий и перегородок и т. д.), и впоследствии горение может возобновиться. Далее необходимо принять по возможности меры по эвакуации людей, тушению пожара и сохранности материальных ценностей [36].

6.5 Вывод по разделу

Значение всех производственных факторов на изучаемом рабочем месте соответствует нормам, которые также были продемонстрированы в данном разделе.

Категория помещения по электробезопасности, согласно ПУЭ, соответствует первому классу – «помещения без повышенной опасности» [37].

Согласно правилам по охране труда при эксплуатации электроустановок персонал должен обладать I группой допуска по электробезопасности. Присвоение группы I по электробезопасности производится путем проведения инструктажа, который должен завершаться проверкой знаний в форме устного опроса и (при необходимости) проверкой приобретенных навыков безопасных способов работы или оказания первой помощи при поражении электрическим током [38].

Категория тяжести труда в офисе по СанПиН 1.2.3685-21 «Гигиенические нормативы и требования к обеспечению безопасности и (или) безвредности для человека факторов среды обитания» относится к категории Iб (работы, производимые сидя, стоя или связанные с ходьбой и сопровождающиеся физическим напряжением).

Рабочее помещение характеризовано категорией «Д» (пониженная пожар опасность) в соответствии с условиями [39].

Согласно постановлению правительства РФ от 31 декабря 2020 года, N 2398 «Об утверждении критериев отнесения объектов, оказывающих негативное воздействие на окружающую среду, к объектам I, II, III и IV категорий» [40] был выбран объект I категории – оказывающих значительное негативное воздействие на окружающую среду и относящихся к областям применения наилучших доступных технологий.

Заключение

В результате построения прогнозной модели по оценки эффективности влияния входных параметров (данных) на добычу нефти трудно извлекаемых запасов нефти методом гидроразрыва пласта были выявлены параметры, которые критическим образом влияют на величину извлечения нефти.

Построенная модель позволяет оценить будущий прогноз добычи от входных параметров конкретного месторождения, что позволяет дать перспективности будущих месторождений и целесообразности их разработки.

Данная модель позволяет сэкономить существенные финансовые средства при разработки будущих нефтяных месторождений.

Список используемых источников

1. Покрепин Б. В. Разработка нефтяных и газовых месторождений //Р.н/ Д. – 2015.
2. Губин Е.И. Методология подготовки больших данных для прогнозного анализа //Современные технологии, экономика и образование: сборник трудов Всероссийской научно-методической конференции, г. Томск, 27-29 декабря 2019 г.—Томск, 2019. – 2019. – С. 27-29.
3. Губин Е.И. Методика подготовки больших данных для прогнозного анализа «Наука и бизнес: пути развития». Выпуск № 3(105). 2020, 2020. – [С. 33-35].
4. Chen M. S., Han J., Yu P. S. Data mining: an overview from a database perspective //IEEE Transactions on Knowledge and data Engineering. – 1996. – Т. 8. – №. 6. – С. 866-883.
5. Roiger R. J. Data mining: a tutorial-based primer. – 2017.
6. Schröer C., Kruse F., Gómez J. M. A systematic literature review on applying CRISP-DM process model //Procedia Computer Science. – 2021. – Т. 181. – С. 526-534.
7. Wirth, Rüdiger, and Jochen Hipp. "CRISP-DM: Towards a standardprocess model for data mining." Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining. Vol. 1. 2000.
8. Shafique U., Qaiser H. A comparative study of data mining process models (KDD, CRISP-DM nd SEMMA) //International Journal of Innovation and Scientific Research. – 2014. – Т. 12. – №. 1. – С. 217-222.
9. Azevedo A., Santos M. F. KDD, SEMMA and CRISP-DM: a parallel overview //IADS-DM. – 2008.
10. Кобзарь А.И. Прикладная математическая статистика. Для инженеров и научных работников. – 2012.

11. Khalid Z. M. et al. Big data analysis for data visualization: A review //International Journal of Science and Business. – 2021. – Т. 5. – №. 2. – С. 64-75.
12. Rubin D. B., Little R. J. A. Statistical analysis with missing data. – John Wiley & Sons, 2019.
13. Enders C. K. Applied missing data analysis. – Guilford Publications, 2022.
14. Aguinis, Herman, Ryan K. Gottfredson, and Harry Joo. "Best-practice recommendations for defining, identifying, and handling outliers." Organizational Research Methods 16.2 (2013): 270-301.
15. Орлова, И. В. "Подход к решению проблемы мультиколлинеарности при анализе влияния факторов на результирующую переменную в моделях регрессии." Фундаментальные исследования 3 (2018): 58-63.
16. Joseph, V. Roshan, and Akhil Vakayil. "SPlit: An optimal method for data splitting." Technometrics 64.2 (2022): 166-176.
17. Hosmer Jr, David W., Stanley Lemeshow, and Rodney X. Sturdivant. Applied logistic regression. Vol. 398. John Wiley & Sons, 2013.
18. Sperandei, Sandro. "Understanding logistic regression analysis." Biochemia medica 24.1 (2014): 12-18.
19. СП 2.4.3648-20. Санитарно-эпидемиологические требования к организациям воспитания и обучения, отдыха и оздоровления детей и молодежи.
20. Трудовой кодекс Российской Федерации от 30.12. 2001 г. № 197– ФЗ (ред. от 01.04.2019 г.). – М., 2015. – 123 с.
21. ГОСТ Р 50923-96. Дисплеи. Рабочее место оператора.
22. ГОСТ 12.2.032-78. Рабочее место при выполнении работ сидя.
23. ГОСТ 12.0.003-2015. Система стандартов безопасности труда (ССБТ). Опасные и вредные производственные факторы. Классификация.
24. СП 52.13330.2016. Естественное и искусственное освещение.

25. СанПиН 1.2.3685-21. Гигиенические нормативы и требования к обеспечению безопасности и (или) безвредности для человека факторов среды обитания.
26. ГОСТ Р 50948-2001. Средства отображения информации индивидуального пользования. Общие эргономические требования и требования безопасности.
27. ГОСТ Р ИСО 9241-5-2009. Эргономические требования к проведению офисных работ с использованием видео дисплейных терминалов.
28. ГОСТ 12.1.003-2014 ССБТ. Шум. Общие требования безопасности.
29. ГОСТ 12.1.038-82 ССБТ. Предельно допустимые значения напряжений прикосновения и токов.
30. Национальный стандарт РФ ГОСТ Р 59057-2020. «Охрана окружающей среды. Земли. Общие требования по рекультивации нарушенных земель»;
31. ПБ 08-624-03. Правила безопасности в нефтяной и газовой промышленности;
32. СанПиН 1.2.3685-21 «Гигиенические нормативы и требования к обеспечению безопасности и (или) безвредности для человека факторов среды обитания»;
33. Федеральный закон от 21.07.1997 N 116-ФЗ (ред. от 29.12.2022) «О промышленной безопасности опасных производственных объектов»;
34. НПБ 105-03 Определение категорий помещений, зданий и наружных установок по взрывопожарной и пожарной опасности.
35. СП 9.13130.2009 Техника пожарная. ОГНЕТУШИТЕЛИ. Требования к эксплуатации.
36. Постановление Правительства РФ от 16 сентября 2020 г. № 1479 Об утверждении Правил противопожарного режима в Российской Федерации.
37. Правила устройства электроустановок. – 7-е изд. – М.: Изд-во НЦ ЭНАС, 1999-2005;

38. Приказ Минтруда России от 15.12.2020 N 903н (ред. от 29.04.2022) "Об утверждении правил по охране труда при эксплуатации электроустановок";
39. СП 12.13130.2009 «Определение категорий помещений, зданий и наружных установок по взрывопожарной и пожарной опасности»;
40. Постановление Правительства РФ от 31.12.2020 №2398 (ред. от 07.10.2021). «Об утверждении критериев отнесения объектов, оказывающих негативное воздействие на окружающую среду, к объектам I, II, III и IV категорий».

Приложение А Наименование раздела на иностранном языке

Раздел ВКР на английском языке

Обучающийся:

Группа	ФИО	Подпись	Дата
8ПМ1И	Еременко Марк Сергеевич		

Консультант-лингвист отделения (НОЦ) школы: ИШИТР

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент Отделения иностранных языков	Уткина Анна Николаевна	к.филос.н		

References and acronyms

objective function: is a real or integer function of several variables to be optimized (minimized or maximized) in order to solve some optimization problem.

big data: term for data sets that cannot be correctly processed by conventional computers or tools due to their volume, speed of arrival to work with such data.

explanatory variable: a factor that has been manipulated in an experiment by a researcher. It is used to determine the change caused in the response variable. An Explanatory Variable is often referred to as an Independent Variable or a Predictor Variable

multicollinearity: a statistical concept where several independent variables in a model are correlated. Two variables are considered perfectly collinear if their correlation coefficient is +/- 1.0. Multicollinearity among independent variables will result in less reliable statistical inferences.

DATA MINING – intellectual data analysis.

CRISP-DM – Cross-Industry Standard Process for Data Mining.

Introduction

Petroleum industry is a very important part of modern economy, which provides industrial sites, transport and civil real estate with energy. However, gradual petering-out of current oil and gas deposits became pressing issue for humanity, so new perspective deposits are need to be found.

Flammable fossil fuels represent complicated mixture of carbohydrates containing non organic contaminants. Carbon hydrates can be found in solid, liquid and gas form depending on composition, temperature and pressure. Under certain circumstances carbon hydrates can be in both liquid and gas forms. Carbon hydrates mixture which are liquid in deposit and on surface at the same time are called oil. Oil contexture is very complicated and may vary not only between different points of extraction but even in the one deposit. Extraction defines physical and chemical properties of oil. There are more than 25000 points of extraction around the globe and about 1300 of them are in Russia. Moreover, oils' properties change during extraction, in transfer or in contact with other liquids or gas [1].

Recent scientific researches and development of electronics allows us to use new, cutting edge methods of geophysical modeling and forecasting such as Big Data analysis.

The goal of this paper is to develop method for location of new carbon hydrates extraction points using model based on Big Data theory. This method is based on complex analysis of geological, geochemical, engineering and historical data of wells' performance.

The rapid development of information technologies, especially in the field of big data, places increased demands on the quality of the initial data. Considering that most of the real data is loosely structured, the question of the "purity" of the latter is critical. The process of collecting and preparing initial data is one of the most time-consuming and complex stages in the analysis of large amounts of data, which sometimes takes up to 80% of the time. The use of statistical methods and modern software can significantly reduce time and financial costs at this stage and improve the efficiency and quality of the final results.

With careful and correct preparation of the initial data, it is possible to increase the predictive power of traditional predictive models by almost 20%.

Results of this work have practical relevance for oil and gas industry because developed data analysis method will allow data specialist and engineers come up with more sufficient decisions and improve competitive power of companies in digital era.

The usage of latest scientific and technical issues is needed to complete goal of the paper.

1 Intellectual Data Analysis

Data Mining in accordance with the term introduced by G. Shapiro in 1989, – generic term for naming a set of methods for detecting data which was previously unknown, non-trivial, and can be practically useful for decision-making in various areas of human activity. Data Mining can be interpreted as in-depth data analysis. Today, companies use Big Data to target customer interactions, streamline operations, prevent fraud threats, etc. Over the past two years, companies such as IBM, Google, Amazon, Uber have created hundreds of jobs for programmers and data scientists. Data mining methods are based on classification, modeling and forecasting methods which use of decision trees, artificial neural networks, genetic algorithms, evolutionary programming, associative memory, fuzzy logic, regression analysis, etc. Such methods require some a priori ideas about the analyzed data [2].

Currently, Data Mining ("intellectual data analysis") is defined as the process of finding hidden patterns in large amounts of data, which are often unstructured and have a variety of formats (in the form of numbers, text, photos, etc.). Most of the data mining time is spent preparing the data: cleaning, aggregating, transforming, and modeling. Another problem is that statistical models are often built on data with a large number of observations or variables. For scalability, statistical methods must be carefully selected and applied. Once researchers have the data, they can start building statistical models. From a number of possible models, the best one is selected as a result of testing. Often, business requirements determine the kind of numerical model, and in many cases, it turns out to be a logistic regression model. This is determined by the fact that it is possible to evaluate the contribution of input variables and make reliable predictive estimates [3].

1.1 CRISP–DM

Nowadays data are becoming more and more important in terms of making management decisions and business development. However, obtained data may be useless if no structured and systematic method of its' analysis and utilization is used.

In this context CRISP–DM can be used as a foundation for development of forecast models and for extraction of useful data [2].

CRISP–DM is a universal standard process developed for processing and analysis of data as well as compositing prediction models. It is consisted of 6 consistent phases (fig. 1)

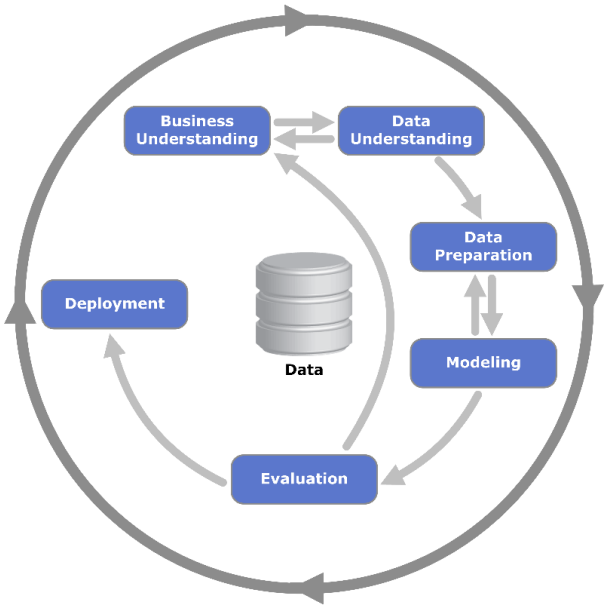

Figure 1 – concept of CRISP–DM

CRISP-DM is based on the following parts:

1. Business understanding – estimation of business goals and evaluation of ways to achieve them in most efficient way.

2. Data understanding – second phase which begins from data acquirement, description and characterization of obtained data and quality control of data.

3. Data preparation – phase in which data specialists prepare data for further use which includes sampling, feature engineering, separation of useless data from shared array, integration and formatting of useful data.

4. Modeling – choice and learning of data models.

5. Evaluation – approbation of algorithm and output, estimation of next steps.

6. Deployment – direct use of model followed by monitoring and receiving of feedback [3].

Advantages of CRISP–DM among other analytic models is flexibility in terms of use in various fields. The methodology provides a systematic approach to working with data, allows you to effectively manage data analysis projects and minimize risks. Because of this, it has gained wide acceptance and recognition in the industry and the scientific community [4].

In this work the phases of CRISP-DM and their application in practice will be considered. Modern approaches and tools related to each phase of the methodology will also be considered.

1.2 SEMMA methodology

The SAS Institute defines data mining as the process of sampling, examining, modifying, modeling and assessing (SEMMA) large amounts of data in order to uncover previously unknown patterns that can be used as a business advantage. The data mining process is applicable across various industries and provides methodologies for such diverse business problems as fraud detection, customer retention, marketing database analysis, market segmentation, risk analysis, similarity analysis, customer satisfaction, bankruptcy forecasting, and loan portfolio analysis [7].

Enterprise Miner contains a set of advanced data mining tools with a common, user-friendly interface that can be used to create and compare multiple models. Statistical tools include clustering, self-organizing maps (Kohonen maps), significant variable selection, decision trees, linear and logistic regression, and neural networks. Data preparation tools include outlier detection, variable transformations, random sampling, and splitting of original datasets for training, testing, and validation. Advanced visualization tools allow you to quickly and easily explore large amounts of data in multivariate histograms and graphically compare simulation results [8].

SAS Enterprise Miner has been designed to support the entire data mining process. The SAS system provides access to relational and heterogeneous data stores. The core SAS language provides high power in data aggregation and transformation. Together, SAS/STAT and Enterprise Miner can support high practical implementation in the numerical simulation of the business process under investigation. Enterprise Miner functions are organized into a well-known logical algorithm as SEMMA (Fig. 2):

• SAMPLE – the formation of an initial dataset for modeling, which must be large enough to contain the necessary information for extraction, and also limited so that it can be effectively used.

• EXPLORE – association detection, visual and interactive statistical analysis, understanding data by detecting expected and unexpected relationships between variables, as well as deviations through data visualization.

• MODIFY – preparing data for analysis (creating additional variables or changing existing variables for analysis, identifying outliers, assessing missing data, changing (modifying) the way input variables are used for analysis, performing cluster analysis, analyzing using Cochran networks, etc.)

• MODEL – application of methods for building and processing data mining models: artificial neural networks, decision trees, regression analysis, etc.

• ASSESS – comparison of simulation results among themselves and with the planned indicators, analysis of the reliability and usefulness of the created models.

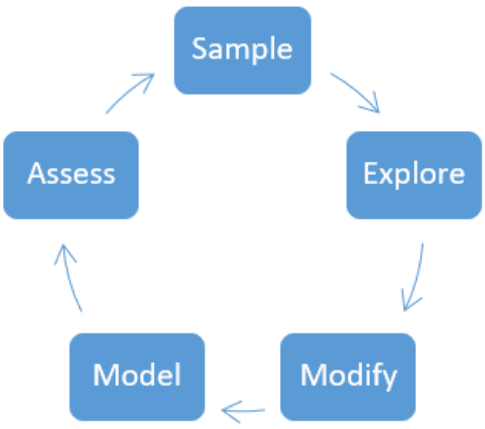

Figure 2 –SEEMA methodology

Within each of the categories, SEMMA Enterprise Miner provides a range of tools to advance the data mining process. The SEMMA methodology is more of a set of recommendations than a set of hard and fast rules.

2. Methodology and preparation of input data

Search and preparation of source material is one of the most complex and time-consuming steps in the process of analyzing a large amount of information, which can take up to 80% of a specialist's working time. Modern statistical techniques and modern software can significantly reduce time, financial and time costs at this stage, as well as improve the quality of the final result [9].

During the initial data preprocessing, the following operations are performed: elimination of errors and anomalies, the study of outliers and duplicate records. An important task of this stage is to identify the presence of multicollinearity between explanatory variables and, if necessary, exclude these variables. The scaling process allows to bring the original data to a single digital format, which significantly increases the accuracy of predictive models [10].

2.1 Input data control

Checking the original data for errors or typographical errors consists in analyzing and evaluating the data in order to identify potential errors in their recording. This process is dedicated to detect random or incorrect values, as well as identify possible typographical errors that can distort the original data and affect the results of the analysis.

Various methods can be used during data control. One common method is analysis of statistical indicators such as mean, median, standard deviation, etc. [11]. Some incorrect values can be detected if they differ significantly from the expected statistics. It is also possible to use data visualization techniques, such as graphing or charting, to visually identify anomalies or unusual values. For example, if the data is presented as a time series, then graphical display can help identify outliers or implausible values [12].

Another approach is to compare the data against known rules or expected values. If the data does not follow predetermined rules or is outside the allowed ranges, then this may indicate the presence of errors or typographical errors.

In general, checking the source data for errors requires careful analysis and expert-based methods. The results of data control can serve as a basis for further data preprocessing and ensuring their reliability and quality before building predictive models.

2.2 Search for missing data values

Identification and processing of missing data are important steps in data analysis. Missing values can occur for a variety of reasons, such as data collection errors, technical problems, or lack of information. However, their presence can seriously distort the results of analysis and forecasting. At the beginning of the process of checking for missing data, we identify missing values in the various variables and columns of the data set. To do this, we check each data record and identify missing or incorrect values.

After identifying the missing data, the next step is to choose the optimal processing strategy. There are several methods for handling missing data, including removing records with missing values, replacing missing values with mean or median values, using interpolation, statistical models, etc. [13].

Choice of strategy for handling missing data depends on the context and features of the dataset. The type of variable (categorical or numerical), the distribution of data, the amount of missing values, as well as possible causes and other factors are taken into account [14].

The purpose of checking and handling missing data is to ensure the completeness and quality of the data in order to eliminate distortions in the analysis and modeling. This allows you to get more accurate and reliable results of research and forecasting based on the available data.

2.3 Search for outliers in input data

Outliers in data is anomaly data values which stand out from data package. Outlier validation is the process of identifying and analyzing observations that

deviate significantly from the expected behavior or distribution of the data. Outliers can occur for a variety of reasons, such as errors in data collection, anomalous events, or natural variations in data.

One method of testing for outliers is to analyze statistical measures such as mean, median, and standard deviation. Outliers may be identified if values differ significantly from the expected statistics or are outside specified limits.

You can also use graphic methods. Outliers in categorical variables can be detected using histograms. The simplest way to define outliers in a numeric variable is to count as an outlier all observations that do not fit into the given quantiles. Graphically, this approach is implemented in the form of box-and-whisker diagrams. These visual representations of the data allow you to identify anomalies and unusual values that could be potential outliers.

Outlier testing helps identify potentially erroneous or anomalous data that can skew the analysis results. With a small number of outliers, you can remove them from the analysis or replace them with an average or mode. With a large number of outliers, they should be separated into a separate sample for analysis, as this may indicate the emergence of a new phenomenon in the data. The use of some transformations (min-max normalization) and discretization can help to cope with outliers in a numeric variable.

References

1. Pokrepin B. V. Razrabotka neftyanyh i gazovyh mestorozhdenij //R.n/D. – 2015.
2. Chen M. S., Han J., Yu P. S. Data mining: an overview from a database perspective //IEEE Transactions on Knowledge and data Engineering. – 1996. – T. 8. – №. 6. – С. 866-883.
3. Roiger R. J. Data mining: a tutorial-based primer. – 2017.
4. Schröer C., Kruse F., Gómez J. M. A systematic literature review on applying CRISP-DM process model //Procedia Computer Science. – 2021. – T. 181. – С. 526-534.
5. Wirth, Rüdiger, and Jochen Hipp. "CRISP-DM: Towards a standardprocess model for data mining." Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining. Vol. 1. 2000.
6. Shafique U., Qaiser H. A comparative study of data mining process models (KDD, CRISP-DM nd SEMMA) //International Journal of Innovation and Scientific Research. – 2014. – Т. 12. – №. 1. – С. 217-222.
7. Azevedo A., Santos M. F. KDD, SEMMA and CRISP-DM: a parallel overview //IADS-DM. – 2008.
8. SAS Enterprise Miner – SEMMA. // SAS Institute URL: https://www.sas.com/en_us/software/enterprise-miner.html (дата обращения: 03.05.2023).
9. Gubin E.I. Metodologiya podgotovki bol'shih dannyh dlya prognoznogo analiza //Sovremennye tekhnologii, ekonomika i obrazovanie: sbornik trudov Vserossijskoj nauchno-metodicheskoj konferencii, g. Tomsk, 27-29 dekabrya 2019 g.—Tomsk, 2019. – 2019. – S. 27-29.
10. Gubin E.I. Metodika podgotovki bol'shih dannyh dlya prognoznogo analiza «Nauka i biznes: puti razvitiya». Vypusk № 3(105). 2020, 2020. – [С. 33-35].

11. Kobzar' A.I. Prikladnaya matematicheskaya statistika. Dlya inzhenerov i nauchnyh rabotnikov. – 2012.

12. Khalid Z. M. et al. Big data analysis for data visualization: A review //International Journal of Science and Business. – 2021. – Т. 5. – №. 2. – С. 64-75.

13. Rubin D. B., Little R. J. A. Statistical analysis with missing data. – John Wiley & Sons, 2019.

14. Enders C. K. Applied missing data analysis. – Guilford Publications, 2022.

Приложение Б

Диаграмма Ганта

Таблица 10 – Календарный план-график проведения НИОКР по теме

	Вид	Исп	Т раб, дн.	Продолжительность														
				Янв.			Фев.			Март			Апр.			Май		
				1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
	Выбор научного руководителя работы	И	8															
	Составление и утверждение темы работы	И	4															
		Р	2															
	Составление календарного плана-графика выполнения работы	И	5															
		Р	1															
	Подбор и изучение литературы по теме работы	И	20															
	Анализ предметной области	И	20															
	Получение данных	И	10															

	Формирование выборок, конструирование признаков, очистка, интеграция и форматирование данных	И	18															
	Выбор, обучение и оценка качества модели.	И	17															
	Согласование выполненной работы с научным руководителем	Р	3															
0	Выполнение других частей работы (финансовый менеджмент, социальная ответственность)	И	7															
1	Подведение итогов, оформление работы	И	5															
		Р	2															

– Руководитель (Р)

– Исполнитель (И)

Итого руководитель потратит на проект около 8 дней, а исполнитель – 114 дней.

Приложение В

Листинг программного кода

```
import pandas as pdimport seaborn as snsimport numpy as npimport matplotlib.p
yplot as pltfrom sklearn import metrics
import sklearn.metricsfrom sklearn.linear_model import LogisticRegression
from sklearn import metrics
from sklearn.model_selection import train_test_split
from sklearn.metrics import confusion_matrix
from sklearn.metrics import    precision_score,    recall_score,    f1_score,
accuracy_score

excel_data = pd.ExcelFile(&amp;apos;Dataset.xlsx&amp;apos;)
df1 = excel_data.parse(&amp;apos;Sheet1&amp;apos;)

df1.dtypes
df1.isnull().sum()

df    =    df1.drop([&amp;apos;Qgrp&amp;apos;,    &amp;apos;Azimut&amp;apos;,    &amp;apos;Pz&amp;apos;,
&amp;apos;Df&amp;apos;, &amp;apos;Dw&amp;apos;], axis=1)
df.boxplot()

for x in [&amp;apos;W&amp;apos;]:
    q75,q25 = np.percentile(df.loc[:,x],[75,25])
    intr_qr = q75-q25

    max = q75+(1.5*intr_qr)
    min = q25-(1.5*intr_qr)

    df.loc[df[x] < min,x] = np.nan
    df.loc[df[x] > max,x] = np.nan

for x in [&amp;apos;K&amp;apos;]:
    q75,q25 = np.percentile(df.loc[:,x],[75,25])
    intr_qr = q75-q25

    max = q75+(1.5*intr_qr)
    min = q25-(1.5*intr_qr)

    df.loc[df[x] < min,x] = np.nan
    df.loc[df[x] > max,x] = np.nan

for x in [&amp;apos;Ht&amp;apos;]:
    q75,q25 = np.percentile(df.loc[:,x],[75,25])
    intr_qr = q75-q25

    max = q75+(1.5*intr_qr)
    min = q25-(1.5*intr_qr)

    df.loc[df[x] < min,x] = np.nan
    df.loc[df[x] > max,x] = np.nan

for x in [&amp;apos;Lt&amp;apos;]:
    q75,q25 = np.percentile(df.loc[:,x],[75,25])
    intr_qr = q75-q25

    max = q75+(1.5*intr_qr)
```

```
    min = q25-(1.5*intr_qr)

    df.loc[df[x] < min,x] = np.nan
    df.loc[df[x] > max,x] = np.nan

df.isnull().sum()
df = df.fillna(df.mean ())

fig, hm = plt.subplots(figsize=(10,10))
hm = sns.heatmap(df.corr(),cbar=True,annot=True)

df = df.drop([&apos;Noil&apos;, &apos;Pplast&apos;, &apos;Mp&apos;], axis=1)

fig, hm = plt.subplots(figsize=(10,10))
hm = sns.heatmap(df.corr(),cbar=True,annot=True)

def categorize_debit(Doil):
    if Doil < 35:
        return 0
    return 1

df[&apos;Doil&apos;] = df.Doil.map(categorize_debit)

print(df)

x                                   =                                   df.loc[:,
(&apos;R&apos;,&apos;W&apos;,&apos;K&apos;,&apos;St&apos;,&apos;Ht&apos;,&apo
s;Lt&apos;,&apos;DeltaP&apos;)]
y = df.loc[:, &apos;Doil&apos;]

X_train, X_test, y_train, y_test = train_test_split(x, y,
                                                    train_size=0.75,
                                                    random_state=42)



model = LogisticRegression().fit(X_train, y_train)
y_pred = model.predict(X_test)

y_pred_proba =  model.predict_proba (X_test)[::,1]
fpr, tpr, _ = metrics.roc_curve (y_test, y_pred_proba)
auc = metrics.roc_auc_score (y_test, y_pred_proba)

#create ROC curve
plt.plot (fpr,tpr,label=" AUC= "+str(round(auc,2)))
plt.ylabel(&apos;True Positive Rate&apos;)
plt.xlabel(&apos;False Positive Rate&apos;)
plt.legend(loc=4)
plt.show()

conf_matrix = confusion_matrix(y_true=y_test, y_pred=y_pred)

fig, ax = plt.subplots(figsize=(5, 5))
ax.matshow(conf_matrix, cmap=plt.cm.Oranges, alpha=0.3)
for i in range(conf_matrix.shape[0]):
    for j in range(conf_matrix.shape[1]):
            ax.text(x=j,    y=i,s=conf_matrix[i,    j],    va=&apos;center&apos;,
ha=&apos;center&apos;, size=&apos;xx-large&apos;)
```

```
plt.xlabel(&apos;Predictions&apos;, fontsize=18)
plt.ylabel(&apos;Actuals&apos;, fontsize=18)
plt.title(&apos;Confusion Matrix&apos;, fontsize=18)
plt.show()

print(&apos;Precision: %.3f&apos; % precision_score(y_test, y_pred))
print(&apos;Recall: %.3f&apos; % recall_score(y_test, y_pred))
       print(&apos;Accuracy: %.3f&apos; % accuracy_score(y_test, y_pred))
```