Министерство науки и высшего образования Российской Федерации
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИТМО»
(Университет ИТМО)

Факультет **Прикладной информатики**

Направление подготовки **09.03.03 Прикладная информатика**

Образовательная программа **Мобильные и сетевые технологии**

К У Р С О В О Й П Р О Е К Т

Тема: «Выбор локации для скважины с помощью ML»

Обучающийся: Данилова Айаана Васильевна, K3141

Санкт-Петербург 2024

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

1 Актуальность темы курсового проекта

Применение технологий машинного обучения (ML) в процессе определения наилучшего местоположения для бурения нефтяных скважин обусловлено рядом весомых причин. На первом месте стоит растущая конкуренция в нефтяной отрасли и стремление к сокращению издержек. Точное определение места для скважины может значительно повысить экономическую отдачу от проектов. Машинное обучение улучшает анализ геологических данных и повышает шансы на успешное обнаружение нефти благодаря более глубокому изучению потенциально перспективных участков.

Следует отметить, что объемы данных в нефтяной сфере велики, что осложняет их обработку традиционными методами. Алгоритмы ML способны эффективно работать с данными о геологии, геофизике и истории разработок, выявляя закономерности, которые могут ускользнуть от внимания человека. Это позволяет специалистам делать более обоснованные выводы и принимать взвешенные решения.

Третий аспект заключается в возможности снижения экологических рисков при выборе мест для бурения. Использование искусственного интеллекта для оценки потенциального воздействия на окружающую среду помогает избегать разработки месторождений в экологически уязвимых зонах, что ведет к более ответственному управлению природными ресурсами.

Последним, но не менее важным фактором, является постоянное развитие технологий и увеличение доступности вычислительных мощностей, что делает применение машинного обучения все более эффективным и доступным. Это обстоятельство подчеркивает актуальность использования ML в поиске и разработке нефтяных месторождений.

Таким образом, использование машинного обучения для определения местоположения нефтяных скважин представляет собой многообещающее направление, которое обеспечивает не только экономическую выгоду для

отрасли, но и способствует защите окружающей среды и более рациональному использованию недр. Это делает тему релевантной для широкого круга заинтересованных сторон, включая нефтяные компании, инвестиционные фирмы, государственные агентства по недропользованию, а также экологов и научных исследователей.

2 Цель проекта

Изучить данные о трех регионах с использованием обученной модели и определить регион, который обеспечит максимальную прибыль.

3 Задачи проекта

- загрузка и проверка данных из трех таблиц;
- корреляционный анализ;
- обучение и валидация моделей для каждого региона;
- расчёт прибыли;
- оценка рисков;
- выбор региона, который принесет наибольшую прибыль.

ОСНОВНАЯ ЧАСТЬ

1 Суть проекта

Перед нефтяной компанией стоит стратегически важная задача: определить оптимальное место для бурения новой скважины, обеспечивающей максимальную прибыль при минимальном риске. Для этого доступны данные разведки трёх регионов, каждый из которых содержит информацию о 10 000 потенциальных месторождений. Для каждого месторождения известны объём запасов нефти (в тысячах баррелей) и качество нефти (характеристики, которые, по условиям задачи, остаются конфиденциальными, но влияют на итоговую прибыль). Задача решается с помощью построения модели машинного обучения [1], способной предсказывать возможную прибыль и риски разработки месторождения.

Шаги для выбора локации:

- в рамках разведки региона производится анализ 10 000 точек, из которых с использованием технологии машинного обучения отбираются 200 самых перспективных для дальнейшей разработки;
- бюджет на разработку скважин в регионе — 10 млрд рублей;
- при текущих ценах каждый баррель сырья приносит доход в размере 450 рублей. Прибыль с каждой единицы продукции составляет 450 тысяч рублей, поскольку объем измеряется в тысячах баррелей;
- после проведения оценки рисков следует оставить только те регионы, где вероятность убытков не превышает 2.5%. Из них выбирается регион с наибольшей средней прибылью;
- характеристики месторождений и подробности контрактов остаются конфиденциальными, не предоставляются для общего доступа.

2 Процессы работы над всем проектом

2.1 Загрузка и валидация данных:

- данные из таблицы были прочитаны и объединены в массив df_all с использованием функции read_excel() из библиотеки pandas;
- было проверено отсутствие пустых строк в таблице с помощью функции dropna();
- было обнаружено, что все метрики, за исключением id, имеют тип данных float64. Следовательно, для построения корреляционной матрицы и обучения модели было принято решение удалить столбец id, поскольку все записи уникальны, и проанализировать данные остальных метрик.

2.2 Корреляционный анализ

- был выполнен корреляционный анализ [2] с помощью функции df_all.corr() и выведена матрица в виде тепловой диаграммы;
- узнали, что имеем умеренную/слабую линейную корреляционную зависимость, следовательно все метрики нам понадобятся для обучения модели.

2.3 Обучение и оценка моделей

- разделение доступного массива данных на две части - обучающую и контрольную, причем 75% данных были отнесены к обучающей выборке, а оставшиеся 25% - к тестовой;
- была обучена модель линейной регрессии;
- были рассчитаны фактические и спрогнозированные моделью значения для данных из тестовой выборки;
- рассчитали корень средней квадратичной ошибки (RMSE). Эта метрика является одним из основных показателей эффективности для модели прогнозирования регрессии [3]. Рассчитывается как квадратный корень из MSE: $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \widehat{y_i})^2}$. Чтобы рассчитать

MSE, надо взять разницу между предсказанными значениями и истинными, возвести её в квадрат и усреднить по всему набору данных [4].

2.4 Расчет прибыли

- анализ прибыльности был выполнен для трех разных наборов данных;
- были вычислены средние значения объемов сырья для каждого из исследуемых регионов, что позволило оценить необходимый объем сырья для обеспечения безубыточности разработки новой скважины;
- был рассчитан средний запас сырья по регионам, на основании чего были отобраны 200 наиболее перспективных месторождений для каждого региона, для которых впоследствии была оценена потенциальная прибыль.

2.5 Оценка рисков

Оценка рисков проводилась с использованием метода Bootstrap [5], начиная с отбора случайной выборки из 500 скважин по каждому региону, из которых затем выбирались 200 с наибольшими запасами нефти. Несмотря на то, что средний объем запасов превышал точку безубыточности в каждом из регионов, для Региона №2 разница составила всего 3 тыс., что существенно повышало риск убытков. С помощью метода Bootstrap были рассчитаны такие показатели, как 95%-й доверительный интервал, средняя прибыль и риск убытков.

2.6 Результат

В результате проведенного анализа был сделан вывод о том, что Регион №2 представляет наибольший интерес для разработки нефтяных скважин, поскольку даже на уровне 0.025 квантиля наблюдается положительное значение, а риск убытков, составляющий 1,6%, удовлетворяет установленному критерию в менее чем 2.5%.

3 Проблема, которая была поставлена передо мной

Передо мной была поставлена проблема подготовки презентации, опираясь на который мы защитим курсовой проект. В рамках этой работы мне необходимо было осуществить следующие пункты:

- планирование презентации – определение целей, изучение аудитории, формирование структуры и логики подачи материала;
- составление сценария – логика, содержание;
- разработка дизайна презентации – определение соотношения текстовой и графической информации, введение анимационных эффектов, цветовая гамма.

4 Решение проблемы

4.1 Планирование презентации

В планировании презентации нашего завершенного проекта мы ставили перед собой задачу не только уложиться в установленное ограничение по времени выступления в 7 минут, но и сохранить высокую информативность презентации, чтобы привлечь внимание аудитории, состоящей из преподавателя, студентов, магистрантов и более опытных специалистов.

Для достижения этой цели мы составили структуру презентации следующим образом: сначала мы представили краткий обзор проекта на первых слайдах, за что отвечала я. Затем каждый слайд содержал информацию о конкретной задаче проекта, способе ее реализации. Мои коллеги по очереди делились информацией о своем вкладе в проект. В конце были подведены итоги.

4.2 Составление сценария

Краткое содержание слайдов представлено в таблице 1.

Таблица 1 - Сценарий презентации

№	Заголовок	Текст	Изображение
1	Выбор локации для скважины с помощью ML		Качественное фото нефтяной вышки
2	Наша команда	Представление членов команды	Фотографии участников
3	Что такое машинное обучение?	Простое объяснение машинного обучения, без сложных терминов.	Простая визуализация концепции

Окончание таблицы 1

4	Инструменты и технологии	Список используемых инструментов и технологий	Логотипы
5	Цели и задачи	Четко сформулированные цели и задачи проекта	
6-10	Задача	Пошаговое объяснение решения, пронумерованная последовательность действий	
11	Вывод	Краткое резюме результатов и дальнейшие перспективы	Ссылка на результат

4.3 Разработка дизайна презентации

В дизайне презентации мы сделали акцент на минималистичном подходе к текстовому наполнению. Это решение основано на психологических особенностях восприятия информации - большие текстовые блоки могут вызывать у аудитории дискомфорт и снижать эффективность восприятия.

Визуальное оформление построено на контрастном сочетании белого шрифта и различных оттенков изумрудно-зеленого цвета, что создает запоминающийся образ, несмотря на сдержанную палитру. Выбор цветовой гаммы неслучаен и перекликается с тематикой курсовой работы, посвященной нефтяной отрасли.

Особую элегантность дизайну придает радиальный градиент фона, дополненный на некоторых слайдах декоративным узором с правой стороны. Такое оформление обеспечивает презентации визуальную привлекательность при сохранении профессионального стиля.

В данном случае отсутствовали какие-либо анимационные эффекты.

Скриншот шестого слайда презентации в итоговом виде представлен на рисунке 1, а десятого слайда презентации на рисунке 2.

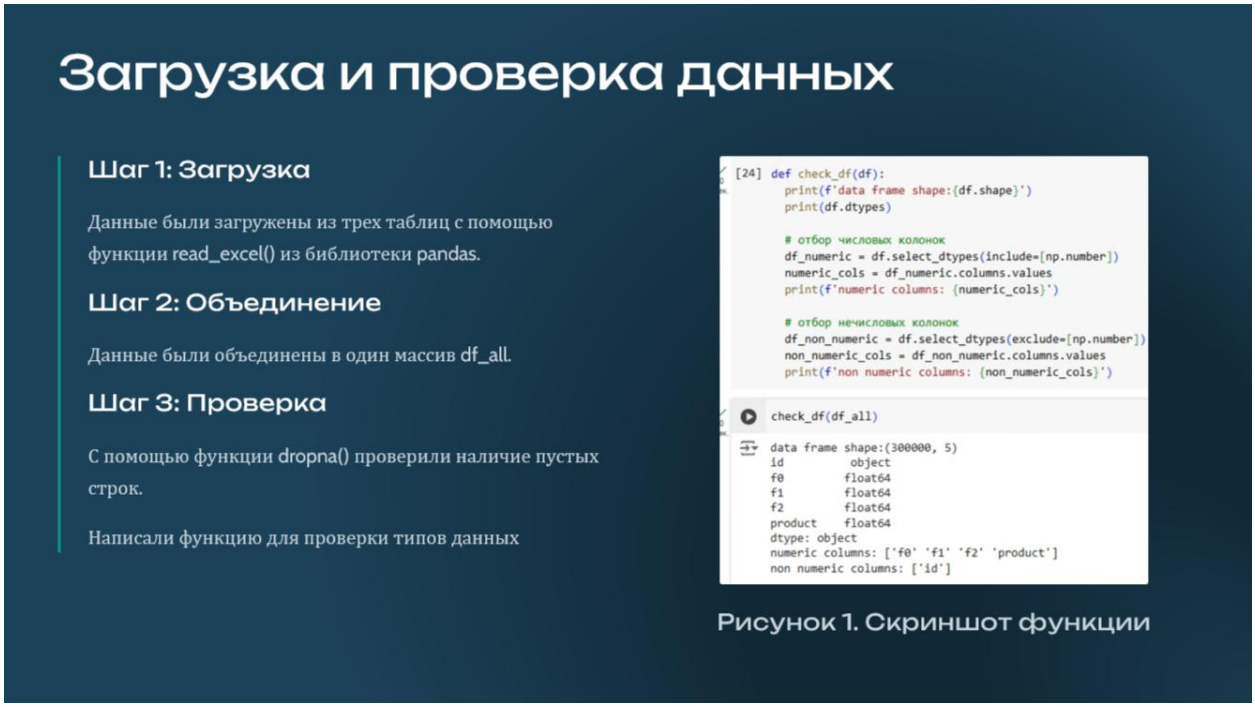

Рис. 1 - Слайд задачи с скриншотом

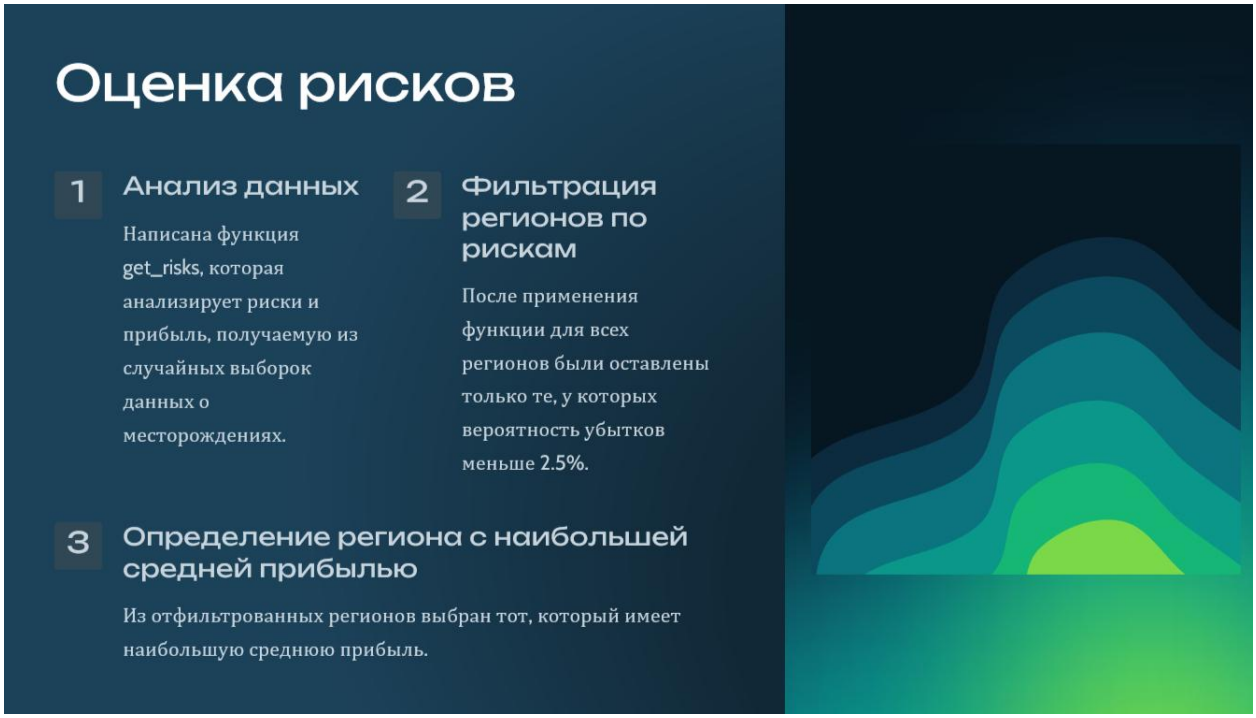

Рис. 2 - Слайд задачи с узором

5 Анализ моей работы

Наша команда получила оценку 4.79 из 5 по критерию продуманной логики изложения и презентации. Этот итог можно считать достаточно хорошим, но всё же не идеальным. Повторный анализ проделанной работы позволил выявить ряд преимуществ и недостатков:

- структура доклада отличается логичностью и последовательностью, охватывая все ключевые аспекты изучаемой темы, начиная от общего введения в область машинного обучения и заканчивая выводами. Такое разделение способствует легкости усвоения материала;
- применение графических элементов, включая концептуальные изображения и скриншоты кода, эффективно способствует демонстрации сложных концепций и повышает наглядность презентации. Скриншоты кода оказываются особенно ценными для иллюстрации технической стороны проекта;
- выбор материалов демонстрирует стремление сделать тему машинного обучения в контексте геологии понятной для студенческой аудитории, что свидетельствует о доступности информации;
- однако было отмечено, что информация о конкретных задачах проекта (слайды 6-10) представлена в слишком краткой форме;
- присутствует риск перегрузки материала кодом. Несмотря на полезность скриншотов кода, их излишнее количество может отвлекать внимание от главной темы доклада;
- отсутствует информация о выбранной модели машинного обучения. Не упоминается, какой именно алгоритм был применен и почему был сделан выбор в его пользу.

6 Взаимодействие с командой

Наша команда состояла из очень отзывчивых и способных ребят, работать с которыми было одним удовольствием. Так как мне необходимо было подготовить презентацию, хотя я и никак не была причастна к практической составляющей, мои коллеги оказались главными источниками информации для меня. Каждый член команды демонстрировал высокую степень компетентности в выполнении своих задач и очень подробно и понятно расписал мне о своей части работы. Также следует упомянуть, что мы активно сотрудничали друг с другом, помогая исправлять ошибки и предлагая решения. Процесс разработки проходил без конфликтов, и мы успешно справились с требованиями технического задания в установленные сроки.

7 Взаимодействие с руководителем

Сразу после формирования команды наш руководитель, Александр, создал общую группу в телеграмме, где предоставил нам техническое задание и спросил насчёт удобного для большинства времени для проведения конференции в зум, а после назначил дату. Во время созвона он провел детальное обсуждение каждого этапа проекта и помог распределить обязанности между участниками. Непосредственно в процессе самой работы Александр направлял нас, оценивая прогресс и предоставляя обратную связь на каждом этапе. Он был открыт к общению и оперативно отвечал на все возникающие вопросы.

8 Оценка работы руководителя

Работа с нашим руководителем оставила исключительно положительные впечатления, никаких явных и неявных недостатков в его работе не было выявлено мной в период нашего сотрудничества. Александр внимательно следил за соблюдением сроков и при необходимости напоминал нам про них, благодаря чему проект был завершен вовремя. Его вклад в нашу работу, по моему мнению, заслуживает высшей оценки.

ЗАКЛЮЧЕНИЕ

1 Оценка выполнения всего проекта

Наша команда успешно завершила каждую задачу, тем самым достигла поставленную цель в установленные временные рамки. В итоге, была разработана и обучена модель, основанная на принципах линейной регрессии. Эта модель способна анализировать информацию о нефтяных месторождениях для прогнозирования объемов возможной добычи нефти, а также оценки потенциальной прибыли и связанных с проектом рисков.

2 Мой вклад в достижение цели

Мною была подготовлена презентация, призванная наглядно представить результаты и выводы курсовой работы. Презентация охватила все ключевые аспекты исследования и, как мне кажется, эффективно донесла необходимую информацию.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1 - Машинное обучение

2 - Корреляционный анализ

3 - Регрессия в машинном обучении

4 - Метрики качества линейных регрессионных моделей

5 - Метод Bootstrap

ПРИЛОЖЕНИЕ

ТЕХНИЧЕСКОЕ ЗАДАНИЕ

1 Название проекта

Выбор локации для скважины с помощью ML.

2 Цель проекта

Определения региона, где добыча принесет наибольшую прибыль.

3 Сроки выполнения

Начало 01 ноября 2024 г..

Конец 20 декабря 2024 г..

4 Руководитель проекта

Иванов Александр Евгеньевич, К4241.

5 Термины и сокращения

- ML - машинное обучение.

6 Требования к проекту

6.1 Технические требования

- для обучения модели подходит только линейная регрессия (остальные — недостаточно предсказуемые);
- при разведке региона исследуют 500 точек, из которых с помощью машинного обучения выбирают 200 лучших для разработки;
- бюджет на разработку скважин в регионе — 10 млрд рублей;
- при нынешних ценах один баррель сырья приносит 450 рублей дохода. Доход с каждой единицы продукта составляет 450 тыс. рублей, поскольку объем указан в тысячах баррелей.;
- после оценки рисков нужно оставить лишь те регионы, в которых вероятность убытков меньше 2.5%. Среди них выбирают регион с наибольшей средней прибылью.;

- данные синтетические: детали контрактов и характеристики месторождений не разглашаются.

6.2 Программные

- Google Colab,
- Python 3.

7 Содержание работы

Содержание работы с ответственными за каждую часть и со сроками выполнения каждой задачи представлено в таблице 1.

Этапы проекта	Сроки выполнения этапов	Ответственный за этап	Вид представления результатов этапа
Разработка технического задания	10 ноября	Иванов Александр	Файл Google doc
Загрузка и проверка данных	14 ноября	Мкртчян Карина	Jupiter notebook
Корреляционный анализ	17 ноября	Мкртчян Карина	Jupiter notebook
Обучение и валидация моделей для каждого региона	1 декабря	Субагио Сатрио	Jupiter notebook
Расчет прибыли	8 декабря	Журбина Марина	Jupiter notebook
Оценка рисков	17 декабря	Усольцева Алина	Jupiter notebook
Защита проекта (сдача отчета и представление доклада с презентацией)	20 декабря	Данилова Айаана	Презентация PowerPoint

Таблица 1 - Этапы проекта и сроки их выполнения