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: Журбина Марина Андреевна, K3139

Санкт-Петербург 2024

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

1 Актуальность темы курсового проекта

В условиях обостряющейся конкуренции и непрерывного роста цен на энергоресурсы, нефтегазовая отрасль находится в постоянном поиске способов повышения эффективности и снижения затрат на всех этапах производственного цикла. Одним из наиболее перспективных направлений оптимизации является применение технологий машинного обучения (ML) при определении оптимального местоположения для бурения нефтяных и газовых скважин.

Во-первых, точное определение местоположения скважины напрямую влияет на экономическую эффективность проекта. Неправильный выбор может привести к значительным финансовым потерям, связанным с низкой продуктивностью скважины или полным отсутствием добычи углеводородов. Использование методов машинного обучения позволяет значительно повысить точность прогнозирования продуктивности скважин, анализируя огромные массивы геолого-геофизических данных. Более глубокий и комплексный анализ данных, предоставляемый ML-алгоритмами, позволяет выявлять скрытые корреляции и закономерности, недоступные традиционным методам обработки информации, что существенно увеличивает вероятность успешного обнаружения и разработки месторождений.

Во-вторых, применение ML в данной области способствует минимизации экологических рисков, связанных с бурением. Определение оптимального местоположения с учетом данных об окружающей среде позволяет избежать бурения в экологически чувствительных зонах, что в свою очередь способствует рациональному использованию природных ресурсов и уменьшает вероятность негативного воздействия на окружающую среду. Анализ потенциального экологического ущерба с помощью ML-моделей позволяет принимать более взвешенные решения, учитывающие не только экономические, но и экологические факторы.

В-третьих, постоянное развитие вычислительных технологий и увеличение доступности мощных вычислительных ресурсов делает применение ML в нефтегазовой отрасли все более эффективным и экономически обоснованным. Современные вычислительные мощности позволяют обрабатывать сложные ML-модели в разумные сроки, что делает данную технологию доступной для широкого круга компаний.

Таким образом, использование машинного обучения для определения местоположения нефтяных скважин представляет собой перспективное направление, сочетающее в себе экономическую эффективность, экологическую ответственность и рациональное использование природных ресурсов. Актуальность данной темы подтверждается интересом к ней со стороны нефтегазовых компаний, инвестиционных фирм, государственных органов, экологических организаций и научного сообщества.

2 Цель проекта

Проанализировать данные о трех регионах, создать и обучить модель на выборке данных, и определить регион, который обеспечит максимальную прибыль.

3 Задачи проекта

- загрузить и проверить данные из трех таблиц,
- выполнить корреляционный анализ,
- выбрать метрики необходимые для обучения моделей,
- обучить и проверить модели для каждого региона,
- рассчитать прибыль,
- оценить риски,
- выбрать регион с наибольшей средней прибылью и наименьшей вероятностью убытков.

ОСНОВНАЯ ЧАСТЬ

1 Суть проекта

Цель проекта заключается в разработке и применении модели машинного обучения [1] для решения стратегически важной задачи выбора оптимального местоположения бурения новой нефтяной скважины. Задача оптимизации направлена на максимизацию потенциальной прибыли и минимизацию рисков, связанных с разработкой месторождения. Для этого будут использованы данные разведки трех регионов, содержащие информацию о 10 000 потенциальных месторождений в каждом из регионов. В качестве входных данных будут использоваться объемы запасов нефти (в тысячах баррелей) и набор конфиденциальных параметров, характеризующих качество нефти и влияющих на конечную прибыльность. Модель должна предсказывать средний запас сырья для каждого из 30 000 потенциальных месторождений, что позволит принять обоснованное решение о местоположении новой скважины.

Шаги для выбора локации:

– из 10 000 потенциальных месторождений в каждом регионе с использованием технологии машинного обучения отбираются 200 самых перспективных для дальнейшей разработки,

– бюджет на разработку скважин в регионе – 10 млрд рублей,

– при нынешних 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 нужно оставить только те регионы, где вероятность убытков не превышает 2.5%. Из них выбирается регион с наибольшей средней прибылью.

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 [2],

– было проверено отсутствие пустых строк в таблице с помощью функции dropna(),

– было обнаружено, что все метрики, за исключением id, имеют тип данных float64. Для построения корреляционной матрицы и обучения модели было принято решение удалить столбец id, поскольку все записи уникальны, затем проанализировать данные остальных метрик.

2.2 Корреляционный анализ

– был выполнен корреляционный анализ [3] 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 тестовые в соотношении 75% к 25%,

– была обучена модель линейной регрессии,

– моделью были спрогнозированы значения для данных из тестовой выборки,

– рассчитали корень средней квадратичной ошибки (RMSE). Эта метрика является одним из основных показателей эффективности для модели прогнозирования регрессии [4]. 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 [5].

2.4 Расчет прибыли

– анализ прибыльности был произведен для трех регионов,

– был вычислен достаточный объем сырья для безубыточной 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

– для демонстрации целесообразности использования техники Bootstrap [6] были взяты сначала случайные 500 скважин по каждому из регионов, из которых были взяты 200 лучших по объему запасов нефти,

– хотя средний объем запасов нефти в каждой из выборок превысил порог безубыточности, маржа прибыли, особенно в Регионе №2 (3 тыс. баррелей), оказалась недостаточно высокой, что существенно увеличивало риск убытков,

– для более точной оценки рисков и потенциальной прибыли был применен Bootstrap-анализ, позволивший определить 95%-й доверительный интервал (на основе 0.025 и 0.975 квантилей), средний объем прибыли и вероятность убытков для каждой выборки.

2.6 Результат

В результате проведенного анализа был сделан вывод, что Регион №2 представляет наибольший интерес для разработки нефтяных скважин. Это обусловлено тем, что даже нижняя граница 95%-ного доверительного интервала (0.025 квантиль) демонстрирует положительную прибыль, а расчетная вероятность убытков (1,6%) ниже допустимого уровня в 2,5%.

3 Проблема, которая была поставлена передо мной

Передо мной была поставлена проблема расчета прибыли по предсказанным данным для трех различных регионов. В рамках этой работы мне необходимо было осуществить следующие задачи:

- подготовка к расчету прибыли,
- выбор скважин с максимальными значениями предсказаний,
- расчет суммы объема сырья для каждого региона,
- расчет прибыли для полученного объема сырья.

4 Решение проблемы

4.1 Подготовка к расчету прибыли

Все необходимые значения для расчета прибыли нужно было сохранить в отдельных переменных.

Затем рассчитать достаточный объем сырья для безубыточной разработки новой скважины и сравнить полученный объем сырья со средним запасом сырья в каждом регионе.

```
[ ] budget = 10 ** 10 # Бюджет на разработку скважин
    count_objects = 200 # Кол-во скважин
    proceeds = 450000 # Доход с каждой единицы продукта
    # Достаточный объём сырья для безубыточной разработки новой скважины
    sufficient_volume_product = round(budget / (count_objects * proceeds), 2)
    # Средний запас сырья в каждом регионе
    mean_pred_region_1 = Y_pred.mean()
    mean_pred_region_2 = Y_pred1.mean()
    mean_pred_region_3 = Y_pred2.mean()
```

```
[ ] print(f'Достаточный объём сырья для безубыточной разработки новой скважины: {sufficient_volume_product}')
    print(f'Средний объём сырья в первом регионе: {mean_pred_region_1}')
    print(f'Средний объём сырья во втором регионе: {mean_pred_region_2}')
    print(f'Средний объём сырья в третьем регионе: {mean_pred_region_3}')
```

```
Достаточный объём сырья для безубыточной разработки новой скважины: 111.11
Средний объём сырья в первом регионе: 92.3987999065777
Средний объём сырья во втором регионе: 68.71287803913762
Средний объём сырья в третьем регионе: 94.77102387765939
```

Как можно заметить, среднее значение объём сырья в каждом регионе меньше достаточного объём для безубыточной разработки скважин. Это означает, что нужно из каждого региона выбрать лучшие точки для разработки новых скважин.

Так же можно заметить, что во втором регионе самый низкий средний показатель, а в третьем самый высокий.

Рисунок 1 – Подготовка к расчету прибыли

4.2 Выбор скважин с максимальными значениями предсказаний

Необходимо было выбрать 200 наиболее перспективных месторождений для каждого региона. Для этого я отсортировала все точки месторождений для каждого региона в обратном порядке (от наиболее богатого скопления нефти к наименее) и взяла 200 первых точек.

```
# Выбор 200 лучших точек для скважины в каждом регионе
best_points_region_1 = sorted(Y_pred, reverse=True)[:200]
best_points_region_2 = sorted(Y_pred1, reverse=True)[:200]
best_points_region_3 = sorted(Y_pred2, reverse=True)[:200]
```

Рисунок 2 – Выбор скважин с максимальными значениями предсказаний

4.3 Расчет суммы объема сырья для каждого региона

Для выполнения этой задачи нужно просуммировать целевое значение объема сырья, соответствующее сделанным предсказаниям в каждом регионе.

```
# Сумма объёма сырья для каждого региона
volume_raw_materials_region_1 = sum(best_points_region_1)
volume_raw_materials_region_2 = sum(best_points_region_2)
volume_raw_materials_region_3 = sum(best_points_region_3)
```

Рисунок 3 – Расчет суммы объема сырья для каждого региона

4.4 Расчет прибыли для полученного объема сырья

Чтобы рассчитать прибыль для полученного объема сырья в каждом из регионов нужно сумму объема сырья в регионе (volume_raw_materials_region_*) умножить на доход с одной единицы продукта (proceeds) и вычесть бюджет (budget), так как прибыль рассчитывается как разница между суммарными доходами и расходами.

```
# Рассчёт прибыли с 200 скважин в каждом регионе
proceeds_region_1 = volume_raw_materials_region_1 * proceeds - budget
proceeds_region_2 = volume_raw_materials_region_2 * proceeds - budget
proceeds_region_3 = volume_raw_materials_region_3 * proceeds - budget
```

Рисунок 4 – Расчет прибыли для полученного объема сырья

5 Анализ моей работы

На основе анализа бюджета и данных по трем регионам, я определила необходимое количество сырья для безубыточной разработки новой скважины, рассчитала средние запасы сырья в каждом из трех представленных регионов, а также потенциальную прибыль для каждого из них.

Благодаря эффективной организации проекта, я успешно выполнила все поставленные задачи в установленный срок.

В ходе работы я успешно применила свои знания в области машинного обучения и анализа данных, что не только обеспечило успешное выполнение моих задач, но и способствовало углублению моих компетенций в этой сфере.

6 Взаимодействие с командой

Взаимодействие в команде строилось на эффективном распределении задач с учетом индивидуальных навыков и опыта участников: некоторые члены команды работали над знакомыми им задачами, другие осваивали новые области.

Наша команда состояла из отзывчивых и способных ребят, работать с которыми было очень приятно, так как все члены команды всегда быстро выходили на связь, активно помогая друг другу в исправлении ошибок, а также вместе преодолевая сложности освоения новой среды разработки (Google Colab). Во время разработки не возникло ни одного конфликта, все задачи, поставленные перед командой, были выполнены в соответствии со сроками, обговоренными заранее.

Я думаю, данный проект способствовал расширению компетенций членов нашей команды в области анализа данных, машинного обучения и практической разработки.

7 Взаимодействие с руководителем

Наш руководитель, Александр, оперативно организовал командную работу: создал чат в Telegram, где предоставил техническое задание и скоординировал время онлайн-встречи в Zoom. Во время встречи мы детально обсудили каждый этап проекта, а руководитель помог распределить обязанности между участниками в соответствии с желаниями и умениями каждого. В течение всей работы Александр осуществлял руководство, отслеживал прогресс, оперативно предоставлял обратную связь на каждом этапе разработки и быстро отвечал на все наши вопросы.

8 Оценка работы руководителя

Работа с нашим руководителем оставила исключительно положительные впечатления, никаких недостатков в его работе я не заметила за период нашего сотрудничества. Александр всегда оперативно отвечал на возникающие вопросы по этапам разработки, помогал и направлял, а также внимательно следил за соблюдением сроков, при необходимости напоминал нам про них, благодаря чему проект был завершен вовремя. Я бы оценила работу нашего руководителя на высший балл.

ЗАКЛЮЧЕНИЕ

1 Оценка выполнения всего проекта

В рамках проекта наша команда успешно разработала и обучила модель линейной регрессии для анализа данных о нефтяных месторождениях, выполнив все задачи в установленные временные рамки. Модель позволяет прогнозировать объемы добычи нефти в потенциальном местоположения бурения новой нефтяной скважины. Также были рассчитаны прибыль и риски для предсказанных моделью значений. В итоге, был выбран наиболее подходящий регион для разработки новой скважины.

2 Мой вклад в достижение цели

Мною были проанализированы предсказания модели и рассчитана потенциальная прибыль для каждого региона. На основе произведенной мной работы можно было переходить к дальнейшему этапу разработки, оценки рисков, после чего можно было определить наилучший регион для разработки новой нефтяной скважины.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Машинное обучение – URL: https://gb.ru/blog/mashinoe-obuchenie
2. Документация библиотеки Pandas – URL: https://pandas.pydata.org/docs/reference/index.html
3. Корреляционный анализ – URL: https://habr.com/ru/companies/nerepetitor/articles/250633/
4. Регрессия в машинном обучении – URL: https://www.geeksforgeeks.org/regression-in-machine-learning/
5. Метрики качества линейных регрессионных моделей – URL: https://loginom.ru/blog/quality-metrics
6. Метод Bootstrap – URL: https://www.ibm.com/docs/ru/spss-statistics/beta?topic=bootstrapping-

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.

Таблица 1 – Этапы проекта и сроки их выполнения

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	Google Colab
Корреляционный анализ	17 ноября	Мкртчян Карина	Google Colab
Обучение и валидация моделей для каждого региона	1 декабря	Субагио Сатрио	Google Colab
Расчет прибыли	8 декабря	Журбина Марина	Google Colab
Оценка рисков	17 декабря	Усольцева Алина	Google Colab
Защита проекта (сдача отчета и представление доклада с презентацией)	20 декабря	Данилова Айаана	Презентация PowerPoint