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 **45.03.04 Интеллектуальные системы в гуманитарной сфере**

Образовательная программа **Языковые модели и искусственный интеллект**

К У Р С О В О Й П Р О Е К Т

Тема: «Выбор локации для скважины с помощью ML»

Обучающийся: Усольцева Алина Дмитриевна, К3161

Санкт-Петербург 2024

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

1 Актуальность проекта

Проект «Выбор локации для скважины с помощью ML» актуален по следующим причинам:

1.1 Повышение экономической эффективности бурения

Использование машинного обучения (ML) [1] позволяет более оптимально выбирать места для бурения скважин, что существенно повышает эффективность, т.е снижает затраты на единицу добытого продукта.

1.2 Снижение рисков бурения

Алгоритмы машинного обучения могут анализировать большие объемы данных, и таким образом выявлять сложные закономерности, которые не были бы очевидны при других, более традиционных методах анализа. Это повышает вероятность нахождения высокоприбыльных месторождений.

1.3 Увеличение скорости принятия решений

Технологии машинного обучения способны значительно ускорить процесс выбора локаций для бурения, а это является важным преимуществом в условиях конкурентной среды и быстро меняющихся рыночных условий.

1.4 Улучшение адаптивности к изменению данных

Использование машинного обучения позволяет быстро адаптироваться к изменениям в данных, например, в геологических условиях, данных о скважинах или изменениях в законодательстве, что является важным аспектом для предприятий в данной области.

2 Цель проекта

С помощью машинного обучения определить, регион в котором добыча нефти принесет наибольшую прибыль.

3 Задачи проекта

1) загрузить и проверить данные,
2) провести корреляционный анализ,
3) обучить и валидировать модели для каждого региона,
4) рассчитать прибыль для каждого региона,
5) провести оценку рисков.

ОСНОВНАЯ ЧАСТЬ

1 Суть проекта

Суть проекта состоит в применении методов машинного обучения для повышения эффективности выбора месторождений для добычи сырья. Он сочетает в себе прогнозирование и экономический анализ, таким образом помогая выбрать наиболее привлекательный регион для инвестиций в бурение скважин с наименьшим риском убытков.

2 Процессы работы над проектом

На этапе загрузки и проверки данные были загружены из трех таблиц, в которые представлены данные о соответствующих трех регионах, с помощью функции read_excel() из библиотеки pandas. Затем данные были объединены в один массив df_all. С помощью функции dropna() проверили наличие пустых строк. Написали функцию check_df (показана на рисунке 1) для проверки типов данных.

```
[24] def check_df(df):
        print(f'data frame shape:{df.shape}')
        print(df.dtypes)

        # отбор числовых колонок
        df_numeric = df.select_dtypes(include=[np.number])
        numeric_cols = df_numeric.columns.values
        print(f'numeric columns: {numeric_cols}')

        # отбор нечисловых колонок
        df_non_numeric = df.select_dtypes(exclude=[np.number])
        non_numeric_cols = df_non_numeric.columns.values
        print(f'non numeric columns: {non_numeric_cols}')
```

```
check_df(df_all)

data frame shape:(300000, 5)
id          object
f0         float64
f1         float64
f2         float64
product    float64
dtype: object
numeric columns: ['f0' 'f1' 'f2' 'product']
non numeric columns: ['id']
```

Рисунок 1 - Функция check_df

Следующим этапом работы стал корреляционный анализ (корреляционная матрица показана на рисунке 2), то есть выявление взаимосвязей двух или нескольких случайных параметров.

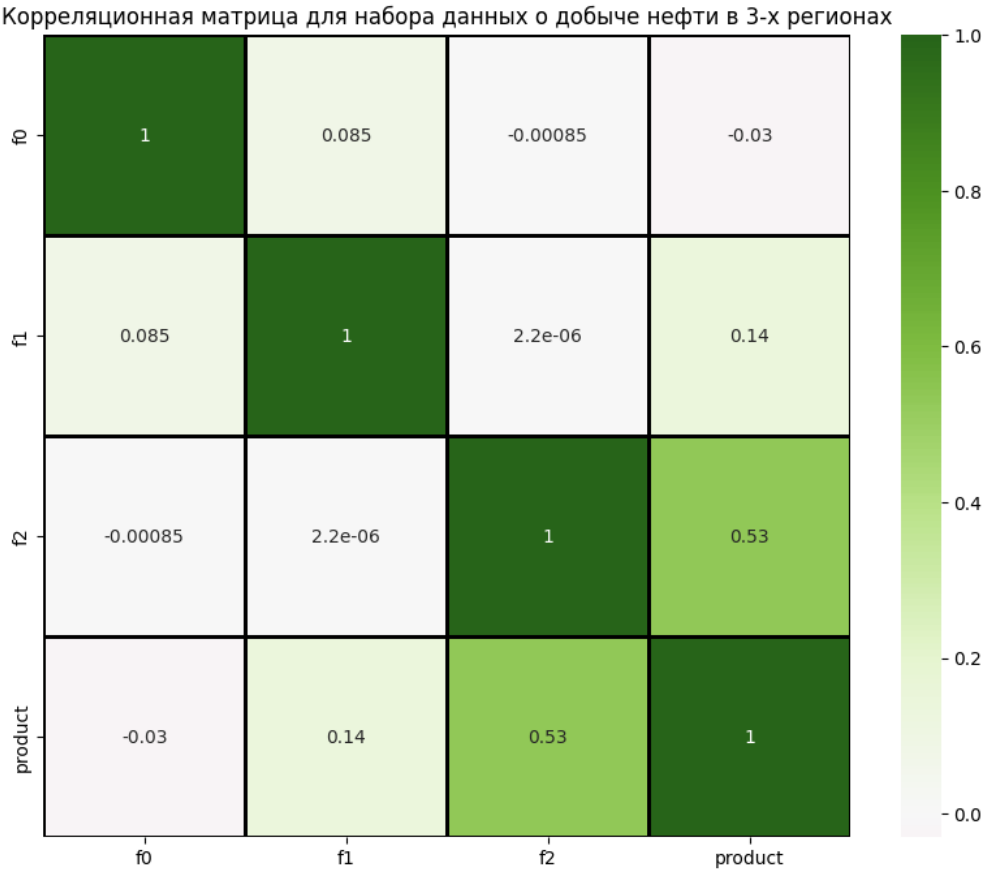

Рисунок 2 - Вывод матрицы корреляции

Данные были разбиты на обучающую и валидационную выборки в соотношении 75:25 и были рассчитаны средние значения для прогнозируемых моделью значений (prediction) и фактических значений из данных проверки (actual). Модель машинного обучения была обучена на обучающих данных с помощью алгоритма линейной регрессии и были сделаны предсказания на валидационной выборке. Метрика RMSE используется для расчета среднего отклонения между прогнозируемыми и фактическими значениями. Таким образом, как результат мы получили выведенные на экран средний запас предсказанного сырья и RMSE модели.

Далее был проведен сравнительный анализ по регионам: модель была проверена на тестовых данных, были рассчитаны RMSE и MEAN, а также сравнены предсказанные значения с фактическими.

Следующим шагом работы является расчет прибыли, в который входит вычисление прибыли, средних значений сырья и среднего запаса сырья.

Сначала все ключевые значения для расчётов были сохранены в отдельных переменных, затем была рассчитана прибыль для 3 различных датафреймов и средние значения объёма сырья в каждом регионе. Это важно, чтобы понять, какой объем сырья необходим для безубыточности новой скважины. В завершение был подсчитан средний запас сырья в каждом регионе. Опираясь на эти данные, мы выбрали 200 лучших точек для бурения скважин в каждом регионе и рассчитали потенциальную прибыль с 200 скважин в каждом регионе.

Функция для расчёта прибыли по выбранным скважинам и предсказаниям модели работает следующим образом (реализация функции показана на рисунке 3):

- выбор скважин с максимальными предсказанными значениями,
- расчет целевого значение объёма сырья, соответствующее этим предсказаниям,
- расчет прибыли для полученного объёма сырья.

Заключительный этапом работы это подсчет рисков и прибыли для каждого региона:

- применена техника Bootstrap с 1000 выборок, чтобы найти распределение прибыли,
- вычислена средняя прибыль, 95%-й доверительный интервал и риск убытков.

```
from numpy.random import RandomState

def get_risks(data):
    size = data.shape[0]
    indexes = pd.Series(range(size))
    profit = []
    risk = 0
    state = RandomState(12345)

    for i in range(1000):
        subsample = indexes.sample(frac=500/size, replace=True, random_state=state)
        profit.append(take_profit(data.loc[subsample, :]))
        if take_profit(data.loc[subsample, :]) < 0:
            risk += 1
    profit = pd.Series(profit)
```

Рисунок 3 - Функция get_risks

На основе этих данных стало возможно предложить регион, наиболее привлекательный для бурения скважин.

3 Суть поставленной задачи

Мое участие в проекте заключалось в оценке рисков. Мне нужно было вычислить риск убытков и 95% доверительный интервал для средней прибыли 200 лучших месторождений и оставить в выборке только те регионы, в которых вероятность убытков меньше 2.5%, среди них выбрать регион с наибольшей средней прибылью. Оценка рисков является важной завершающей частью проекта, так как определяет точность модели и ее применимость.

4 Методика решения

Для начала я написала функцию get_risks, которая использует технологию бутстрэп с 1000 выборок [2], то есть многократно берет случайные выборки.

В цикле будет 1000 итераций, на каждой из которых случайно выбирается подсэмпл (случайная выборка при помощи RandomState из модуля numpy.random [3]) из исходных данных, рассчитывается прибыль для выбранного подсэмпла и это значение добавляется в список. Функция возвращает как прибыль, так и риски [4], то есть вероятность убытков.

Затем я реализовала функцию get_risks для данных каждого из трех регионов. Выводятся риски и 95% доверительный интервал [5] как диапазон от 2,5-его до 97,5-его процентиля.

В нулевом регионе мы получили вероятность убытков, равную 6%, в первом - менее 2%, а во втором 7% (как показано на рисунке 4). То есть, нашему требованию рисков менее 2,5 процентов соответствует только регион 1.

```
# Посчитаем основные параметры для трёх регионов
print('1-й регион: ')
get_risks(result_prediction0)
print()
print('2-ой регион: ')
get_risks(result_prediction1)
print()
print('3-й регион: ')
get_risks(result_prediction2)
```

```
1-й регион:
Средняя прибыль: 406278783
95-ый доверительный интервал: от -117742136 до 911737051
Риск (вероятность убытка): 0.067

2-ой регион:
Средняя прибыль: 432624132
95-ый доверительный интервал: от 16846175 до 815972527
Риск (вероятность убытка): 0.019

3-й регион:
Средняя прибыль: 377362192
95-ый доверительный интервал: от -170780417 до 901772131
Риск (вероятность убытка): 0.074
```

Рисунок 4 - Результат оценки рисков

Первый регион в целом очень привлекателен для инвестиций в бурение скважин, так как реалистичный прогноз прибыли в нем - 0,5 миллиардов рублей, а оптимистичный - 0,9 миллиардов рублей.

Таким образом, результатом нашей работы стал вывод о максимальной выгодности региона 1, данные о возможной прибыли и вероятности убытков в этом регионе.

5 Анализ работы

Я успешно справилась с поставленной передо мной задачей, мне удавалось работать над проектом планомерно. Я разбила поставленную задачу на несколько подзадач, начиная с разработки функции get_risks и заканчивая анализом данных по каждому региону и формулировкой итоговых выводов. Я старалась следовать намеченному плану и укладываться в сроки, отведенные на выполнение части проекта, за которую я ответственна.

Несмотря на общую планомерность моего рабочего процесса, возникали некоторые сложности, например, на этапе разработки функции get_risks мне потребовалось дополнительное время, чтобы разобраться с тонкостями реализации метода Bootstrap и правильно интерпретировать полученные результаты. Кроме того, при анализе данных по регионам иногда возникали вопросы, требующие дополнительной проверки и уточнения, что также немного замедляло процесс. Эти небольшие задержки были связаны с необходимостью глубокого понимания принципов работы, а не просто с выполнением задания "по шаблону".

За время работы над проектом я углубила свои знания в области машинного обучения, в частности, я на практике изучила метод Bootstrap и его применение для оценки рисков. Я научилась применять этот метод для анализа реальных данных и получила опыт в интерпретации полученных результатов, что является важным навыком для анализа данных.

6 Взаимодействие с командой

Данный проект подразумевает последовательную работу, то есть один участник выполняет свою задачу, и только затем к работе приступает следующий. По ходу реализации проекта мы с командой общались в общем чате, делились результатами своей работы, получали обратную связь как от руководителя, так и от коллег по команде.

Перед защитой проекта я поделилась подробным описанием выполненной мной задачи, чтобы помочь участнику, составляющему презентацию, точнее описать процесс работы над проектом.

7 Взаимодействие с руководителем проекта

После того, как команда была сформирована, руководитель сразу связался с каждым из участников и добавил в общий чат. Перед началом работы руководитель организовал общий созвон для команды, на котором у нас была возможность познакомиться, выбрать интересующую нас задачу в проекте и получить ответы на любые возникшие вопросы.

В дальнейшем к Александру можно было обращаться для уточнения требований к своим задачам, подсказок в исправлении ошибок в коде, помощи в работе в новой для многих из нас среде Google Colab, а также с любыми другими вопросами. Руководитель своевременно и подробно отвечал.

8 Оценка работы руководителя

Считаю работу руководителя хорошей. Из положительного хочу отметить умение Александра распределить задачи и организовать слаженную работу команды. Также Александр быстро реагировал на сообщения, был готов помочь с любыми возникшими трудностями.

ЗАКЛЮЧЕНИЕ

1 Оценка выполнения проекта

Наша команда достигла поставленной цели в указанные сроки, каждый ее член успешно справился со своими задачами. Проект по созданию прогнозной модели для нефтегазовой отрасли был успешно реализован. Наша модель, основанная на линейной регрессии, точно анализирует данные со скважин и позволяет прогнозировать объемы добываемой нефти в рассматриваемом регионе. Модель также оценивает потенциальную прибыль и риски для каждого месторождения, способствуя оптимизации решений и эффективному управлению ресурсами.

2 Личный вклад в достижение цели

Я взяла на себя ключевую роль в анализе рисков и обосновании выбора региона в рамках данного проекта. Основным моим вкладом стала разработка и реализация функции get_risks, основанной на методе Bootstrap, с 1000 выборок. Эта функция позволяет многократно пересемплировать исходные данные, а это, в свою очередь, дает возможность оценить риски убытков для каждого региона.

СПИСОК ИСПОЛЬЗУЕМЫХ ИСТОЧНИКОВ

1) “Машинное обучение”, А.А. Миронов,
2) Machine Learning: What is Bootstrapping?,
3) Документация numpy.random,
4) Методы оценки экономических рисков,
5) Понятие доверительного интервала,
6) ГОСТ 7.32-2017

ПРИЛОЖЕНИЕ

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	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