

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

КУРСОВОЙ ПРОЕКТ

Тема: «Выбор локации для скважины с помощью ML»

Обучающийся: Мкртчян Карина Геворговна, К3141

Санкт-Петербург 2024

СОДЕРЖАНИЕ

ВВЕДЕНИЕ	3
1 Актуальность темы курсового проекта	3
2 Цель проекта	4
3 Задачи проекта	4
ОСНОВНАЯ ЧАСТЬ	4
1 Суть проекта	5
2 Процессы работы над всем проектом	6
3 Проблема, которая была поставлена передо мной	9
4 Решение проблемы	10
5 Анализ моей работы	13
6 Взаимодействие с командой	14
7 Взаимодействие с руководителем	15
8 Оценка работы руководителя	16
ЗАКЛЮЧЕНИЕ	17
1 Оценка выполнения всего проекта	17
2 Мой вклад в достижение цели	17
СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ	18
ПРИЛОЖЕНИЕ	19

ВВЕДЕНИЕ

1 Актуальность темы курсового проекта

Актуальность темы выбора локации для нефтяной скважины с помощью машинного обучения (ML) обусловлена несколькими ключевыми факторами.

Во-первых, нефтяная индустрия сталкивается с растущей конкуренцией и необходимостью оптимизации затрат. Выбор правильной локации для бурения скважины может существенно влиять на экономическую эффективность проектов. Использование машинного обучения позволяет более точно анализировать геологические данные и выявлять потенциально богатые ресурсы, повышая вероятность успешного бурения.

Во-вторых, машинное обучение предоставляет инструменты для обработки больших объемов данных, свойственных нефтяной отрасли. Данные о геологии, геофизике, истории бурения и другие факторы могут быть сложно интерпретировать традиционными методами. Алгоритмы ML способны находить закономерности и корреляции, что помогает геологам и инженерам принимать более обоснованные решения.

В-третьих, применение ML в выборке локации для нефтяных скважин способствует снижению экологических рисков. Искусственный интеллект может помочь в оценке воздействия на окружающую среду и избежать размещения скважин в уязвимых экосистемах, тем самым способствуя более устойчивому использованию природных ресурсов.

Наконец, с развитием технологий и увеличением доступности вычислительных мощностей применение машинного обучения становится всё более доступным и эффективным. Эта тема остается актуальной, учитывая необходимость постоянного совершенствования методов и подходов в поиске и разработке природных ресурсов.

Таким образом, выбор локации для нефтяной скважины с помощью машинного обучения не только отвечает на экономические потребности

отрасли, но и способствует экологической безопасности и рациональному использованию ресурсов.

Использование машинного обучения для выбора локации бурения скважин позволяет значительно улучшить эффективность и рентабельность операций, а также минимизировать экологические риски, что делает данную тему весьма актуальной для многих игроков на рынке, например для

- нефтяных компаний (минимизация затрат на разведку и бурение);
- инвестиционных фирм (обоснованный выбор локаций для бурения может повысить их уверенность в инвестициях);
- государственных органов, а именно федерального агентства по недропользованию (использование модели для оценки потенциальных экологических последствий и рационального использования ресурсов);
- экологов и исследователей (использование результатов моделей для оценки воздействия на экосистемы и разработки стратегий минимизации ущерба).

2 Цель проекта

Проанализировать данные о трех регионах с помощью обученной модели и выбрать регион, который принесет наибольшую 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

Нужно решить, где бурить новую скважину для нефтяной компании. Предоставлены пробы нефти в трёх регионах: в каждом 10 000 месторождений, где измерили качество нефти и объём её запасов. Необходимо построить модель машинного обучения, которая поможет определить регион, где добыча принесет наибольшую прибыль. Необходимо проанализировать возможную прибыль и риски техникой Bootstrap.

Шаги для выбора локации:

- при разведке региона исследуют 10 000 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.

2 Процессы работы над всем проектом

2.1 Загрузка и валидация данных:

- с помощью функции из библиотеки pandas - `read_excel()` - считали данные из таблиц и объединили их в массив `df_all`;
- с помощью функции `dropna()` - проверили отсутствие пустых строк в таблице;
- узнали, что все метрики, кроме `id` имеют тип данных `float64`, соответственно для построения корреляционной матрицы и обучения модели мы можем удалить столбец `id`, так как все записи уникальны и проанализировать данные остальных метрик.

2.2 Корреляционный анализ

- выполнили корреляционный анализ с помощью функции `df_all.corr()` и вывели матрицу в виде тепловой диаграммы;
- мы имеем умеренную/слабую линейную корреляционную зависимость, следовательно все метрики нам понадобятся для обучения модели.

2.3 Обучение и оценка моделей

- разделили данных: данные были разделены на обучающие и тестовые в соотношении 75% к 25%;
- обучили модель линейной регрессии;
- рассчитать 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. Рассчитывается как квадратный корень из MSE: $RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$. Чтобы рассчитать MSE, надо взять разницу между предсказанными

значениями и истинными, возвести её в квадрат и усреднить по всему набору данных [1].

2.4 Расчет прибыли

- расчет прибыли для 3 различных датафреймов;
- вычисление средних значений сырья: мы рассчитали средние значения объема сырья в каждом регионе. Это важно, чтобы понять, какой объем сырья необходим для безубыточной разработки новой скважины;
- вычисление среднего запаса сырья: мы подсчитали средний запас сырья в каждом регионе. Опираясь на эти данные, мы выбрали 200 лучших точек для скважины в каждом регионе и рассчитали потенциальную прибыль с 200 скважин в каждом регионе.

2.5 Оценка рисков

- для демонстрации целесообразности использования техники бутстреп были взяты сначала случайные 500 скважин по каждому из регионов, из которых были взяты 200 лучших по объема запасов нефти;
- расчет среднего показателя объема запасов хоть и превосходил точку безубыточности для каждого из регионов, однако, например, для Региона №2 лишь на 3 тыс., ввиду чего риск убытков значительно повышается;
- были рассмотрены показатели с помощью техники бутстреп, рассчитав при этом следующие показатели: 95%-й доверительный интервал (0.025 и 0.975 - квантили), среднее объем прибыли, а также риск убытков.

2.6 Результат

В конечном итоге, проанализировав значения необходимых показателей, можно сделать вывод о том, что наиболее привлекательным регионом для разработки скважин нефти будет Регион №2, поскольку даже 0.025-квантиль имеет положительное значение, а также риск убытков удовлетворяет необходимому требованию - менее 2.5% - и составляет 1,6%.

3 Проблема, которая была поставлена передо мной

Передо мной была поставлена проблема анализа данных и выявление зависимости между каждой парой метрик. Я должна была выполнить две данные задачи:

- загрузить и проверить данные из трех таблиц;
- провести корреляционный анализ и построить корреляционную матрицу на основе всех данных.

4 Решение проблемы

4.1 Загрузка и проверка данных

- импортируем нужные библиотеки;

```
import numpy as np
import seaborn as sns
import pandas as pd
import matplotlib.pyplot as plt
```

Рис. 1 - Импорт библиотек в Google Colab'е

- с помощью функции из библиотеки pandas - read_excel() - считаем данные из таблиц и объединим их в массив df_all. С помощью функции dropna() - проверили отсутствие пустых строк в таблице[2].;

```
df0 = pd.read_excel('/content/geo_data_0.xlsx')
df0 = df0.dropna()
```

```
df1 = pd.read_excel('/content/geo_data_1.xlsx')
df1 = df1.dropna()
```

```
df2 = pd.read_excel('/content/geo_data_2.xlsx')
df2 = df2.dropna()
```

```
df_all = pd.concat([df0, df1, df2])
```

Рис. 2 - Валидация данных

- напомним функцию для проверки типов данных;

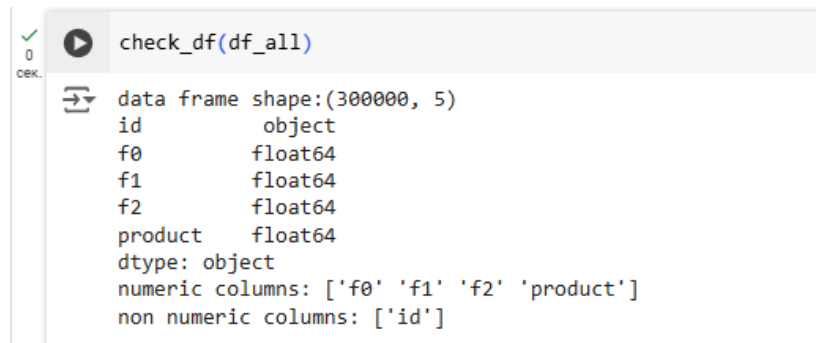
```
[24] def check_df(df):
      print(f'data frame shape:{df.shape}')
      print(df.dtypes)

      # 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}')
```

Рис. 3 - Проверка данных

- применим функцию к датасету.



```
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']
```

Рис. 4 - Результат проверки данных

Все столбцы, кроме id имеют тип float64. Следовательно мы можем проанализировать эти данные: построить корреляционную матрицу, а затем обучить модель, так как столбец id нам для этого не понадобится.

4.2 Корреляционный анализ

Напишем функцию, которая рассчитывает коэффициенты корреляции для каждой пары метрик с помощью функции corr() из библиотеки Pandas и выводит эту таблицу в виде тепловой диаграммы с помощью библиотек Matplotlib [3] и Seaborn [4].

```
def correlation_matrix(df):
    print(df.head())
    df = df[['f0', 'f1', 'f2', 'product']]

    corr_matrix = df.corr()
    print(corr_matrix)

    plt.figure(figsize=(10, 8))

    sns.heatmap(corr_matrix, annot=True, cmap='PiYG', center=0, linewidths=1, linecolor='black')

    plt.title('Корреляционная матрица для набора данных о добыче нефти в 3-х регионах')
    plt.show()
```

Рис. 5 - Функция для выявления линейной корреляционной зависимости

Применим данную функцию к датасету и построили корреляционную матрицу для данных из трех таблиц [5].

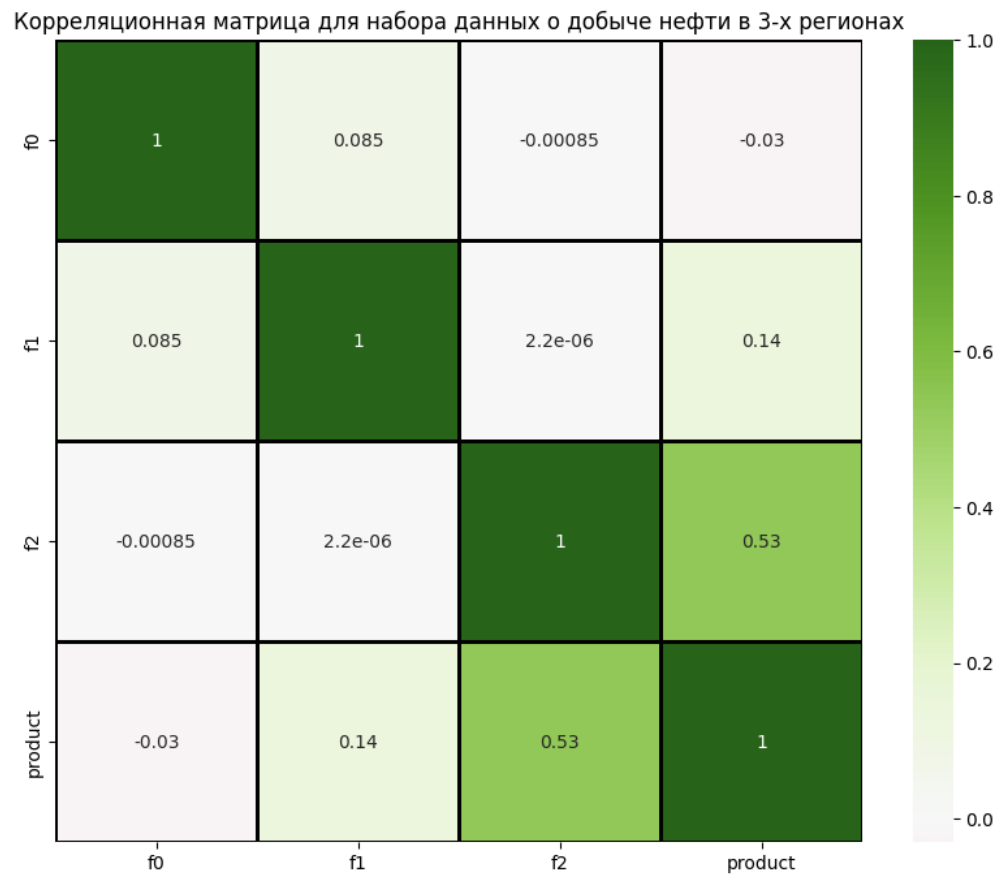


Рис. 6 - Вывод матрицы с коэффициентами корреляции с помощью библиотеки Seaborn

5 Анализ моей работы

В результате проделанной мной работы, я выяснила, что данные подходят для анализа и обучения модели, так как метрики имеют тип `float64` и в таблице отсутствуют пустые строки. Каждая строка уникальна, поэтому мы можем не использовать столбец `id`, который тип `object`.

Также с помощью корреляционного анализа я выяснила, что между метриками нет сильной связи, значит при обучении мы можем использовать данные из всех столбцов, и переобучения возникнуть не должно.

Мне удалось работать планомерно и я выполнила свои задачи намного раньше указанного срока.

Во время выполнения курсового проекта я применила свои имеющиеся знания в аналитике данных и вспомнила некоторые забытые нюансы.

6 Взаимодействие с командой

Члены нашей команды быстро распределили обязанности: кто-то взялся за те задачи, которые ему/ей уже знакомы, кто-то столкнулся с новыми для себя проблемами. Несмотря на это, на мой взгляд, все члены команды отлично справились со своей задачей. Мы помогали друг другу исправить ошибки и недочеты, подсказывали, как работать в новой для некоторых среде - Google Colab. Во время разработки у нас не возникло конфликтов и мы выполнили все, что написано в техническом задании в соответствии со сроками.

Я думаю, что данный курсовой проект был полезен для нашей команды, мы познакомились с аналитикой данных и машинным обучением, а также побыли в роли разработчиков.

7 Взаимодействие с руководителем

Руководитель сразу же связался с нами после формирования команды. Александр отправил нам техническое задание, а затем мы подробно обсудили каждый этап выполнения проекта на созвоне и распределили обязанности. Во время выполнения проекта руководитель направлял нас и оценивал полученный на каждом этапе результат. Александр отвечал на все наши вопросы и вовремя давал обратную связь.

8 Оценка работы руководителя

Мне понравилось работать с нашим руководителем, никаких минусов я для себя не заметила. Александр следил за сроками выполнения каждой задачи, поэтому мы закончили даже раньше указанной даты. Я бы оценила его работу на максимальный балл.

ЗАКЛЮЧЕНИЕ

1 Оценка выполнения всего проекта

Цель проекта была достигнута нашей командой в указанные сроки, все поставленные задачи были выполнены участниками, ответственными за них. В результате мы получили обученную модель на основе линейной регрессии, которая анализирует данные о месторождении и прогнозирует количество нефти, которое возможно добыть, а также прибыль и риски.

2 Мой вклад в достижение цели

Я подготовила данные для анализа и обучения, рассчитала корреляционную матрицу, по результатам которой наглядно видно, что коэффициент корреляции Пирсона относится к умеренной или слабой зависимости, значит при использовании всех метрик модель не должна переобучиться.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

- 1 - [Метрики качества линейных регрессионных моделей](#)
- 2 - [Документация библиотеки Pandas](#)
- 3 - [Документация библиотеки Matplotlib](#)
- 4 - [Документация библиотеки Seaborn](#)
- 5 - [Расчет коэффициента корреляции Пирсона](#)

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, K4241.

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

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