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: Субагио Сатрио Брахманторо Ади, K3140

Санкт-Петербург 2024

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

1 Актуальность темы курсового проекта

В современную эпоху нефтяная промышленность сталкивается со все более сложными проблемами, сопровождающимися ростом потребностей в энергоресурсах и жесткой глобальной конкуренцией. Одной из главных задач является поиск более эффективных и действенных способов разведки новых нефтяных месторождений, учитывая высокие затраты на процесс бурения, а также риск возможных потерь. Поэтому использование передовых технологий, таких как машинное обучение, становится все более актуальным для решения этой проблемы.

Технология машинного обучения позволяет анализировать геологические и геофизические данные в больших масштабах с большей точностью, чем традиционные методы. Такие данные, как качество нефти, объем запасов и история предыдущего бурения, могут быть обработаны алгоритмами, способными выявить скрытые закономерности и дать рекомендации по выбору оптимальных мест для бурения. Это очень важно для нефтяных компаний в их стремлении минимизировать риски и оптимизировать прибыль.

Кроме того, применение моделей на основе машинного обучения помогает снизить негативное воздействие на окружающую среду. Благодаря более точным прогнозам компании могут избегать бурения в экологически уязвимых районах, поддерживая тем самым принцип более ответственной и устойчивой эксплуатации природных ресурсов.

Помимо экономических и экологических выгод, актуальность этой темы заключается также в ее потенциале для развития инноваций в энергетическом секторе. Использование передовых технологий при принятии стратегических решений может стать первым шагом к дальнейшей цифровизации процесса разведки и добычи нефти. Это открывает перед компаниями широкие

возможности для повышения конкурентоспособности на международном рынке.

Таким образом, выбор места для бурения нефти с использованием машинного обучения - это не только решение для повышения операционной эффективности, но и стратегический шаг, способствующий устойчивому развитию энергетической отрасли. Данная тема весьма актуальна для широкого круга заинтересованных сторон, включая нефтяные компании, инвесторов, правительства, а также научное сообщество, все из которых должны сыграть важную роль в обеспечении более эффективного управления энергетическими ресурсами в будущем.

2 Цель проекта

Главная цель данного проекта — разработка модели машинного обучения, способной анализировать данные разведки из трех регионов и выбирать область с наибольшей прибылью.

3 Задачи проекта

- Загрузите, очистите и проверьте данные 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

В этом проекте основной задачей было разработать модель машинного обучения, которая могла бы давать рекомендации по выбору оптимальных

мест для бурения нефтяных скважин. Данные о геологоразведке в трех регионах были проанализированы с помощью различных статистических методов и алгоритмов прогнозирования, чтобы определить регионы с наибольшим потенциалом прибыли и наименьшим риском.

В каждом регионе имеется подробная информация о 10 000 потенциальных точках бурения, включая объемы запасов нефти и другие геологические характеристики. Конечная цель - выбрать 200 лучших точек из каждого региона для дальнейшей разработки. Бюджет, выделенный на проект, составляет 10 миллиардов рублей, а целевая выручка с барреля нефти - 450 рублей.

Шаги для выбора локации:

- Анализ исследовательских данных из 10 000 точек в каждом регионе с помощью алгоритмов машинного обучения.;
- По результатам анализа отобраны 200 лучших точек для дальнейшей разработки;
- Рассчитали потенциальную рентабельность каждого региона с учетом прогнозируемого объема запасов нефти и рыночных цен;
- Выделили бюджет в размере 10 миллиардов рублей на развитие выбранных точек;
- Для оценки риска каждого региона использовали метод бутстрепа;
- Отсеяли регионы с риском потерь более 2,5 %;
- Выбрали участки с наибольшей средней прибылью и наименьшим риском потерь в качестве рекомендуемых для бурения.

2 Процессы работы над всем проектом

2.1 Сбор и проверка данных

- Собрали данные разведки из трех регионов с помощью функции read_excel() библиотеки Pandas и объединили их в массив df_all.
- Проверили данные с помощью функции dropna(), чтобы убедиться, что в таблице нет пустых строк.

- Определили, что все метрики, кроме столбца ID, имеют тип данных float64. Поэтому столбец ID был удален, так как все записи уникальны, а для анализа и обучения модели использовались только остальные метрики.

2.2 Корреляционный анализ

- Проведите корреляционный анализ между различными геологическими параметрами с помощью функции corr(), чтобы понять взаимосвязь между объемом запасов нефти и другими характеристиками.
- Выявлены параметры, оказывающие значительное влияние на результаты прогнозирования, и эти результаты использованы в качестве основы для обучения модели.

2.3 Обучение модели

- Разделите данные на два набора: обучающие (75 %) и тестовые (25 %).
- Обучите линейную регрессионную модель на обучающих данных и оцените ее эффективность с помощью показателей среднеквадратичной ошибки (RMSE).
- Используя результаты оценки, оптимизируйте модель, прежде чем использовать ее для расчета прибыли.

2.4 Расчет прибыли

- Рассчитали потенциальную прибыль каждого региона на основе прогнозируемого объема запасов нефти и рыночной цены за баррель.
- Рассчитали точку безубыточности, чтобы убедиться, что разработка скважины не принесет убытков, и оценили среднюю прибыль по 200 лучшим точкам.

2.5 Оценка рисков

- Выполнили оценку риска с помощью метода бутстрапа, взяв случайные выборки из данных для измерения изменчивости прогнозируемых результатов.

- Рассчитайте 95%-ный доверительный интервал и оцените вероятность убытка для каждого региона.
- Отфильтруйте регионы с риском потерь более 2,5 % и выберите регионы с наибольшей средней прибылью и наименьшим риском в качестве основных кандидатов на развитие.

2.6 Результат

На основе проведённого анализа было установлено, что Регион №2 является наиболее перспективным для разработки нефтяных месторождений, так как даже при уровне квантиля 0,025 сохраняется положительное значение, а уровень риска убытков в 1,6% соответствует установленному порогу, не превышающему 2,5%.

3 Проблема, которая была поставлена передо мной

Передо мной была поставлена задача обучить и проверить модель для каждого региона. Для выполнения задачи я должен был реализовать следующее:

- Использовать данные геологоразведки из каждого региона и разделить их на обучающие и тестовые данные в пропорции 75 % для обучения и 25 % для тестирования.
- Построить линейную регрессионную модель на обучающих данных для прогнозирования объема запасов нефти в каждом регионе.

- Использование метрики Root Mean Square Error (RMSE) для оценки точности модели на тестовых данных.
- Сравнение результатов прогнозирования модели с фактическими данными для оценки степени соответствия модели.

4 Решение проблемы

4.1 разделите обучающие и тестовые данные

- импортируем нужные библиотеки;

```
[ ]  from sklearn.model_selection import train_test_split
     from sklearn.linear_model import LinearRegression
     from sklearn.metrics import mean_squared_error
```

Рис. 1 - Импорт библиотек в Google Colab’e

- использование функции train_test_split из библиотеки sklearn.model_selection - разделение данных из каждого региона на 75 % обучающих данных и 25 % тестовых или проверочных данных

```
#разделение данных
X = df0[['f0', 'f1', 'f2']]
Y = df0[['product']]

# Разделите данные на обучающие (75%) и проверочные (25%).
X_train, X_val, Y_train, Y_val = train_test_split(X,Y, test_size=0.25, random_state=42)

print(X_train.shape, X_val.shape)
```

```
(75000, 3) (25000, 3)
```

Рис. 2 - разделите обучающие и проверочные данные

4.2 обучение модели

Создадим модель линейной регрессии с использованием библиотеки scikit-learn. Обучим её на тренировочных данных, сделаем предсказания на основе данных для валидации.

```
model = LinearRegression()

# Обучить модель с помощью обучающих данных
model.fit(X_train, Y_train)

#Делайте прогнозы на основе данных проверки
Y_pred = model.predict(X_val)

#Сохраняйте прогнозные и фактические значения
Y_pred = np.ravel(Y_pred)
Y_val = np.ravel(Y_val)
result_prediction0=pd.DataFrame({'Prediction':Y_pred, 'Actual':Y_val})
```

Рис. 3 - разделите обучающие и проверочные данные

4.3 оцените точность модели с помощью RMSE

Для оценки качества предсказаний модели линейной регрессии были рассчитаны средние значения предсказанных и фактических значений, а также метрика RMSE (корень из среднеквадратичной ошибки), которая показывает среднее отклонение предсказанных значений от фактических.

```
#Вычислите средние прогнозируемые и фактические значения
mean_pred = result_prediction0['Prediction'].mean()
mean_actual = result_prediction0['Actual'].mean()

#Оценка модели с помощью RMSE
rmse = np.sqrt(mean_squared_error(Y_val, Y_pred))
```

```
[ ] print("mean prediction:", mean_pred)
    print("mean actual:", mean_actual)
    print(f"RMSE: {rmse}")
```

```
mean prediction: 92.3987999065777
mean actual: 92.32595637084387
RMSE: 37.756600350261685
```

Рис. 4 - Вычисление средних значений и RMSE для оценки качества модели

Для измерения точности регрессионной модели используется метрика RMSE, которая показывает среднюю ошибку предсказания модели по отношению к фактическому значению. Чем меньше значение RMSE, тем выше качество модели. Средние значения прогноза и фактических данных практически совпадают, что говорит о хорошей оценке модели.

4.4 Сравнение прогнозов модели с фактическими данными

Для более наглядной оценки результатов работы модели линейной регрессии была выведена таблица, содержащая первые 5 строк предсказанных и фактических значений на основе данных валидации. Это позволяет визуально оценить качество предсказаний модели.

```
[ ] print("Прогнозируемые и фактические значения по данным валидации:")
    print(result_prediction0.head())
```

```
Прогнозируемые и фактические значения по данным валидации:
   Prediction      Actual
0  101.901017  122.073350
1   78.217774   48.738540
2  115.266901  131.338088
3  105.618618   88.327757
4   97.980185   36.959266
```

Рис. 5 - сравнение прогнозируемых и фактических значений

5 Анализ моей работы

В результате своей работы я построил линейную регрессионную модель для прогнозирования значений на основе обучающих данных.

После построения линейной регрессионной модели я оценил ее эффективность с помощью метрики RMSE и среднего значения предсказанных и фактических значений. Итоговая таблица с предсказанными

и фактическими значениями показывает, что модель предсказывает довольно точно в большинстве случаев. Значение RMSE составляет 37,76, что указывает на приемлемый уровень точности данной модели.

Работа над проектом была организована эффективно, поэтому я смог выполнить все этапы до установленного срока. В процессе работы я применил свои знания в области обучения моделей машинного обучения.

Таким образом, проделанная работа не только позволила мне успешно выполнить поставленные задачи, но и углубила мое понимание алгоритмов машинного обучения.

6 Взаимодействие с командой

В этом проекте наша команда сразу же взяла инициативу в свои руки, распределив обязанности поровну. Одни члены команды выбирали задания, которые соответствовали их навыкам, а другие использовали возможность попробовать что-то новое и взяться за решение задач, с которыми они раньше не сталкивались. Это создало продуктивную атмосферу обучения, где каждый участник старался внести свой посильный вклад.

На протяжении всего процесса ключевую роль играла командная работа. Мы делились информацией, помогали решать проблемы и обсуждали наиболее эффективные подходы. Использование Google Colab в качестве инструмента для совместной работы привело некоторых членов команды к новому способу работы, но благодаря взаимной поддержке все быстро адаптировались.

Этот проект принес много пользы нам как команде. Помимо обогащения наших знаний в области машинного обучения, мы также научились эффективно сотрудничать при разработке проектов. Этот опыт заложил прочную основу для того, чтобы в будущем мы могли с большей уверенностью браться за подобные проекты.

7 Взаимодействие с руководителем

После формирования команды руководитель проекта провел совещание, на котором распределил задачи в соответствии со способностями каждого.
Александр прислал рамочную программу, на которую команда могла ссылаться, чтобы все процессы были понятны.
В ходе проекта менеджер контролировал и оценивал результаты, помогал нам, если мы путались в реализации проекта, кроме того, менеджер отвечал на все вопросы и исправлял проблемы, которые мы не могли устранить самостоятельно.

8 Оценка работы руководителя

Мне очень понравилось работать с нашим менеджером. Особенно для меня, иностранного студента, он помогает мне во многих вопросах и хорошо оценивает мою работу. Кроме того, менеджер всегда напоминает нам о необходимости выполнять работу в установленные сроки, чтобы мы могли хорошо завершить проект. Я ставлю менеджеру максимальный балл.

ЗАКЛЮЧЕНИЕ

1 Оценка выполнения всего проекта

Задачи, поставленные перед нашей командой, были успешно выполнены в установленный срок. Каждый участник команды справился со своей частью работы, благодаря чему мы достигли основной цели проекта — создание модели для анализа данных о месторождениях. Итогом работы стала обученная модель линейной регрессии, способная прогнозировать объем возможной добычи нефти, оценивать потенциальную прибыль и определять уровень рисков, связанных с разработкой месторождений. Процесс разработки модели включал в себя не только обучение алгоритма, но и тщательный анализ исходных данных, подготовку признаков и оценку качества полученных предсказаний. На каждом этапе участники команды взаимодействовали друг с другом, что позволило минимизировать ошибки и улучшить конечный результат. Таким образом, проект помог нам не только достичь поставленной цели, но и получить ценный опыт в анализе данных и построении предсказательных моделей.

2 Мой вклад в достижение цели

Я разделил обучающие и проверочные или тестовые данные, провел обучение с помощью модели машинного обучения с линейной регрессией, использовал матрицу RMSE для оценки результатов обучения и сравнил предсказанные данные с фактическими.

В результате для каждого региона было получено сравнение данных, предсказанных линейной регрессией, и исходных данных, что позволило легко определить прибыль и величину риска для каждого региона.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1 - Машинное Обучение

2 - Библиотека Sklearn

3 - Среднеквадратичной Ошибки (RMSE)

4 - Алгоритм Линейной Регрессии

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