**Урок 6.** Семинар. Создание проекта машинного обучения

1. Подумайте о реальной проблеме, которую, по вашему мнению, можно решить с помощью машинного обучения. Четко определите эту проблему и поставьте SMART-цели для ее решения. Помните, что это должна быть проблема, в которой машинное обучение может помочь сделать прогноз, классифицировать данные или дать рекомендации.
2. Определите потенциальные источники данных, которые могут быть использованы для решения вашей проблемы. Какие данные вам понадобятся? Где вы можете их получить? Напишите краткое резюме вашей стратегии сбора данных.
3. Даже если у вас нет реальных данных для анализа, представьте, что они у вас есть. Как бы вы стали исследовать и понимать эти данные? Что бы вы искали? Напишите краткое описание вашей стратегии анализа исследовательских данных.
4. Вспомните различные модели машинного обучения, которые мы обсуждали, в частности модель "Дерево решений". Какой тип модели машинного обучения будет наиболее подходящим для вашей проблемы и почему? Объясните свой выбор.
5. Напишите краткое описание того, как вы будете обучать и оценивать выбранную вами модель машинного обучения, используя воображаемый набор данных.
6. Наконец, подумайте о том, как вы могли бы развернуть свою обученную модель. Где она будет использоваться? Кто будет ее использовать? Какую пользу она принесет им? Напишите краткое резюме вашей стратегии развертывания.

Помните, что целью этого задания является не создание реальной модели машинного обучения (пока!), а применение концепций, которые вы изучили на этом семинаре, к реальной проблеме.

**1. Определение проблемы и SMART-цели**

В газодобывающей отрасли часто возникают непредвиденные остановки оборудования, что приводит к значительным финансовым потерям и снижению производительности. Причины отказов могут быть связаны с техническими неисправностями, изменениями в составе газа, ухудшением состояния трубопровода и другими факторами. Необходима система предсказания вероятности выхода оборудования из строя, чтобы вовремя проводить профилактическое обслуживание.

SMART-цели:

* S (Specific): Разработать модель машинного обучения для прогнозирования вероятности выхода из строя ключевых компонентов оборудования на основе исторических данных за 30 дней до их возможного возникновения;
* M (Measurable): Достигнуть точности прогнозирования не менее 85% на тестовом наборе данных;
* A (Achievable): Использовать данные о предыдущих поломках, времени работы оборудования, условиях эксплуатации и техобслуживании, чтобы построить предсказательную модель;
* R (Relevant): Модель поможет снизить количество непредвиденных остановок и сократит незапланированные простои, поможет оптимизировать график технического обслуживания и снизит затраты на ремонты;
* T (Time-bound): Создать и протестировать рабочий прототип модели в течение 6 месяцев.

**2. Источники и стратегия сбора данных**

Источники данных:

* Исторические данные о поломках оборудования (дата, время, тип поломки);
* Данные о техническом обслуживании (плановые и внеплановые ремонты, замены деталей);
* Условия эксплуатации (температура, давление, влажность, нагрузки на оборудование);
* Данные о времени работы оборудования (количество часов работы, циклы включения/выключения).

Источники данных:

* Внутренние базы данных компаний, занимающихся газодобычей;
* SCADA-системы и данные с датчиков и систем мониторинга оборудования на месторождениях;
* Отчеты о техническом обслуживании и ремонтах;
* Открытые данные о погодных условиях и геологии региона.

Стратегия сбора данных:

* Сотрудничество с газодобывающими компаниями для получения анонимизированных исторических данных;
* Интеграция данных из SCADA-систем в хранилище данных;
* Установка датчиков на оборудование для сбора данных в реальном времени;
* Регулярная очистка и фильтрация данных;
* Синхронизация информации о ремонтах и отказах с эксплуатационными параметрами;
* Проведение опросов среди технического персонала для получения дополнительных данных о поломках и обслуживании.

**3. Стратегия анализа исследовательских данных**

Исследование данных:

* Провести первичный анализ данных для выявления пропусков и выбросов;
* Использовать визуализацию данных (гистограммы, диаграммы рассеяния) для изучения взаимосвязей между переменными (например, время работы и вероятность поломки);
* Применить корреляционный анализ для определения факторов, наиболее связанных с поломками оборудования;
* Провести анализ временных рядов эксплуатационных параметров;
* Визуализация трендов временных рядов;
* Анализ распределения параметров при нормальной и аварийной работе.

Что искать:

* Определить ключевые предикторы отказов;
* Выявить корреляции между отказами и параметрами работы скважин;
* Идентификация ключевых факторов, способствующих поломкам (например, условия эксплуатации, частота техобслуживания);
* Выявление закономерностей в данных, которые могут помочь в прогнозировании;
* Проверка наличия сезонных закономерностей;
* Применение методов снижения размерности (PCA) для выбора информативных признаков.

**4. Выбор модели машинного обучения**

Для данной задачи наиболее подходящей моделью будет дерево решений случайный лес или градиентный бустинг.

Случайный лес обеспечивает высокую точность и устойчивость к переобучению благодаря ансамблированию деревьев решений, что делает его эффективным для сложных задач с множеством факторов.

Градиентный бустинг позволяет улучшить точность модели за счет последовательного обучения слабых моделей, что также может быть полезно для выявления сложных зависимостей в данных.

Эти модели позволяют анализировать важность признаков.

**5. Обучение и оценка модели**

Обучение модели:

* Разделение данных на тренировочную (70%), валидационную (15%) и тестовую (15%) выборки;
* Обучение модели на тренировочной выборке, используя кросс-валидацию для повышения надежности и оптимизации гиперпараметров. Подбор гиперпараметров (глубина дерева, число деревьев, минимальное количество объектов в листе);
* Преобразование данных (нормализация, заполнение пропущенных значений).

Оценка модели:

* Использовать метрики оценки, такие как точность, полнота, F1-мера и ROC-AUC, для оценки качества модели на тестовой выборке;
* Оценка по метрикам:
* Accuracy (доля верных предсказаний);
* Precision, Recall, F1-score (важны при дисбалансе классов);
* ROC-AUC (показывает качество предсказаний);
* Провести анализ ошибок для понимания, где модель ошибается, и какие улучшения можно внести.

Ожидаемый результат:

* Модель с точностью не менее 85%;
* Выявление ключевых факторов, влияющих на отказы.

**6. Стратегия развертывания**

Модель будет интегрирована в систему управления активами газодобывающей компании в центрах мониторинга газодобывающих предприятий и встраиваться в системы SCADA для автоматического предупреждения отказов

Технический персонал и менеджеры будут использовать модель для оценки вероятности выхода оборудования из строя и планирования профилактического обслуживания.

Это поможет снизить количество непредвиденных остановок оборудования, что приведет к увеличению производительности и снижению затрат на ремонты.

Поможет оптимизировать график технического обслуживания, что позволит более эффективно распределять ресурсы.

Стратегия развертывания следующая:

* Обучение пользователей работе с моделью и интерпретации ее выводов;
* Интеграция модели в существующую IT-инфраструктуру (SCADA, ERP);
* Разработка дашбордов для визуализации предсказаний;
* Создание интерфейса для ввода данных и получения прогнозов;
* Регулярное обновление модели на основе новых данных и отзывов пользователей для повышения ее эффективности;
* Настройка системы уведомлений о возможных отказах.

Будущие пользователи это операторы и инженеры по эксплуатации, а также руководство для планирования ремонтов.

Модель может принести следующую пользу:

* Снижение незапланированных простоев;
* Уменьшение затрат на аварийные ремонты;
* Повышение безопасности эксплуатации.

**Вывод:**

Проект машинного обучения поможет прогнозировать отказы скважин, что позволит оптимизировать обслуживание оборудования, снизить затраты и повысить надежность добычи газа.