

Примеры BigData для ML

Case 1: Рекомендательная Система для Электронной Коммерции

Цели:

- Увеличение продаж через персонализированные рекомендации
- Улучшение удовлетворенности пользователей путём предоставления более релевантного контента

Конвейер Big Data:

1. **Сбор данных:** Информация о поведении пользователя, покупках и интеракциях с товарами собирается в реальном времени.
2. **Предварительная обработка и очистка:** Удаление дубликатов, заполнение пропущенных значений, трансформация данных.
3. **Хранение:** Данные хранятся в распределенной системе хранения, например, Hadoop HDFS или Amazon S3.
4. **Обработка и агрегация:** Используются технологии как Apache Spark для агрегации и преобразования данных.
5. **Feature Engineering:** Создание новых переменных, таких как средний чек, частота покупок и т.д.
6. **Обучение модели ML:** Используются алгоритмы машинного обучения для рекомендаций.
7. **Оценка и Тюнинг:** Модель тестируется на отложенных данных; гиперпараметры оптимизируются.
8. **Деплой в Производство:** Модель интегрируется в систему электронной коммерции.
9. **Мониторинг и Обновление:** Постоянный мониторинг эффективности модели; переобучение при необходимости.

Case 2: Прогнозирование оттока клиентов в телекоммуникационной компании

Цели:

- Снижение оттока клиентов
- Идентификация "рисковых" клиентов для проведения целевых маркетинговых кампаний

Конвейер Big Data:

1. **Сбор данных:** Данные о платежах, использовании услуг, жалобах и обратной связи собираются из различных источников.
2. **Предварительная обработка и очистка:** Форматирование данных, удаление аномалий.
3. **Хранение:** Данные хранятся в Big Data-решении, как Google BigQuery.
4. **ETL (Extract, Transform, Load):** Данные преобразуются и агрегируются.
5. **Feature Engineering:** Создание переменных, которые могут быть индикаторами оттока.
6. **Обучение модели ML:** Используется алгоритмы для классификации.
7. **Оценка и Тюнинг:** Модель тестируется на отложенных данных; гиперпараметры оптимизируются.
8. **Деплой в Производство:** Модель интегрируется с CRM-системой компании для автоматической идентификации "рисковых" клиентов.
9. **Мониторинг и Обновление:** Регулярное обновление модели для адаптации к новым данным и тенденциям.

В обоих случаях, конвейеры Big Data и машинного обучения спроектированы так, чтобы достичь конкретных бизнес-целей, и они включают в себя этапы от сбора и обработки данных до обучения, оценки и развертывания моделей машинного обучения.