**Урок 3. Создание проекта машинного обучения**

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

Проблема: Предприятие хочет автоматизировать классификацию отзывов клиентов на положительные и отрицательные, чтобы быстро реагировать на негативные отзывы и улучшать качество обслуживания.

SMART-цель: За 3 месяца разработать модель машинного обучения, которая сможет точно классифицировать отзывы клиентов с точностью не менее 85%, и внедрить ее в систему обработки отзывов для автоматической сортировки.

1. ***Источники данных и стратегия сбора данных***

Потенциальные источники данных: отзывы клиентов, собранные через онлайн-формы, социальные сети, электронную почту и внутренние базы данных компании.

Необходимые данные: текст отзывов, дата, источник, оценка (если есть), метки "положительный" или "отрицательный" (если доступны).

Стратегия сбора данных: интеграция с системами сбора отзывов, автоматический экспорт текстовых данных в базу данных, а также ручной сбор дополнительных отзывов при необходимости. Важно обеспечить качество данных и их репрезентативность.

1. ***Стратегия анализа исследовательских данных***

При наличии собранных данных я бы начал с их предварительной обработки: очистки текста (удаление знаков препинания, стоп-слов), лемматизации или стемминга. Затем я бы провел анализ распределения отзывов по классам, выявил бы наиболее часто встречающиеся слова и фразы для каждой категории.

Далее я бы использовал методы визуализации (например, облака слов) для понимания ключевых характеристик отзывов. После этого — разделил бы данные на обучающую и тестовую выборки для последующего моделирования.

1. ***Выбор модели машинного обучения***

Для задачи классификации текста наиболее подходящей моделью будет модель "Дерево решений" или более сложные алгоритмы, такие как случайный лес или градиентный бустинг. Однако для начальной оценки можно использовать дерево решений из-за его интерпретируемости.

Обоснование выбора: дерево решений хорошо работает с категориальными данными и позволяет понять важность признаков. Для более высокой точности можно рассмотреть использование методов на основе градиентного бустинга или нейронных сетей.

**Обучение и оценка модели:**

1. Обучение: разделю данные на обучающую и тестовую выборки (например, 80/20). На обучающей выборке обучу модель с использованием векторизации текста (TF-IDF). Настрою гиперпараметры через кросс-валидацию.
2. Оценка: проверю точность, полноту, F1-мера на тестовой выборке. Проанализирую ошибки модели для улучшения признаков или выбора модели.

***Стратегия*** **развертывания**:

Модель будет интегрирована в систему обработки отзывов компании через API или встроенное приложение. Она будет автоматически классифицировать новые отзывы в реальном времени или по расписанию.

Использователи: сотрудники службы поддержки и менеджеры по качеству обслуживания.

Польза: ускорение обработки отзывов, своевременное реагирование на негативные комментарии, улучшение клиентского опыта и повышение репутации компании.