Тестовое задание на позицию Data Scientist

Задача

Предложить решение для автоматической классификации объявлений и товаров по категориям на сайте.

Описание задания

Представьте, что вы работаете над платформой для размещения объявлений (например, интернет-магазин, маркетплейс или сервис частных объявлений). На сайте есть множество объявлений, которые вручную добавляются пользователями. Задача — разработать подход к автоматической классификации объявлений или товаров по категориям, чтобы улучшить процесс модерации и поиска.

Кандидат может использовать:

- Открытые данные с существующих сайтов (например, данные из Kaggle или парсинг данных сайтов с объявлениями, если это возможно).
- Придумать собственную структуру каталога и тестовые данные.

Основные цели:

1. **Разработать подход к классификации товаров и услуг**:
    - Выбрать алгоритм машинного обучения, который подходит для данной задачи (база должна быть на основе текста, но так же можно использовать изображения или комбинации).
    - Обосновать выбор модели.
    - Реализовать базовое решение для классификации.
2. **Рассмотреть возможные проблемы и подводные камни**:
    - Определить потенциальные проблемы классификации, такие как:
        - Перекрытие категорий.
        - Многоуровневая структура каталога.
        - Ошибочные данные, вводимые пользователями.

    - Предложить пути их решения.
3. **Документировать предложенное решение**:
    - Описать процесс разработки модели.
    - Подготовить рекомендации по улучшению решения в будущем.

Требования к выполнению:

1. **Данные**:
    - Кандидат может использовать готовый набор данных (например, с платформы Kaggle) или создать собственный набор данных (например, сгенерировать текстовые описания товаров и категорий).
2. **Реализация**:
    - Разработка кода (Python или другой удобный язык программирования).
    - Использование фреймворков машинного обучения (например, scikit-learn, TensorFlow, PyTorch и др.).
    - Рекомендуется, но не обязательно: провести базовый анализ данных (EDA).
3. **Результаты**:
    - Прототип модели классификации.
    - Описание структуры данных.
    - Метрики качества модели (например, accuracy, F1-score, precision, recall).
    - Описание архитектуры модели и гипотез, используемых при разработке.
4. **Документация**:
    - Отчет с описанием предложенного решения, его преимуществ и недостатков.
    - Презентация возможных улучшений для повышения точности модели в будущем.

Результаты выполнения задания

В результате выполнения задания ожидается:

- Код с решением задачи (например, ноутбук Jupyter с описанием шагов).
- Описание данных, методов, результатов и выводов.
- Документ (или краткая презентация) с обоснованием выбора решений и предложениями по их улучшению.

Рекомендуемые инструменты:

- Python (numpy, pandas, scikit-learn, matplotlib, seaborn).
- Любые библиотеки NLP или работы с текстом (например, spaCy, NLTK, Hugging Face).
- Фреймворки машинного обучения (TensorFlow, PyTorch).

Критерии оценки:

1. Качество предложенного решения.
2. Обоснованность выбора алгоритма и подхода.
3. Анализ возможных проблем и их решение.
4. Документированность и структурированность предоставленного материала.

Дополнительно:

Если кандидат хочет продемонстрировать более сложное решение (например, использование нейронных сетей или мультимодальных данных), это будет дополнительным плюсом, но не является обязательным.