Tourch (model\_team) age

1) придумать метрику качества выборки

2) придумать алгоритм для повышения качества выборки

Данные табличные

Есть данные и для хорошей модели и для плохой

Данные по каждому партнеру есть по итерациям (1 раз в квартал, например)

Доступные фичи: 700-800 признаков

Ограничения по архитектурам ? GPU &

Спринт: составить план

1) метрика

2) обучение модели

3) алгоритм улучшения данных

4) пайплайн для всего этого

Уточнить задачу:

- Сформулировать входные данные (что дано?)

- Что показывает датасет?

- Что на выходе: модель(регрессия/классификация), есть готовая или нужно обучать

- мы должны реализовать только пайплайн или мы должны реализовать сервис целиком и продумать его инфраструктуру?

- нужно исправить выборку или по метрикам определить можно ли для выборки чтото сделать

Команда подготовлена ко встрече. Представлены:

- Драфт пайплайна работы над проектом - Драфт инструментов и технологий, необходимых для реализации проекта - Драфт метрик и критериев успеха (команды и проекта) - План возможных рисков и как их избежать

План

1 Составить литобзор

2 Разработка метрики для оценки качества датасета (на какой части датасета оцениваем метрику) (Python, pandas, numpy, spicy, matplotlib, seaborn, sklearn):

1) Провести EDA данных.

* Оценить формат и типы данных
* Оценить полноту
* оценить отсутствующие значения
* проверить дубликаты
* выбросы
* распределение признаков по классам
* оценить корреляцию между признаками, между признаками и целевой переменной
* 2) составить возможный список метрик (разный для разных типов признаков)
* 3) проанализировать метрики и вывести пороговые значения
* 4) вывести общую метрику

3. Обучение модели (Python, pandas, numpy, scipy, sklearn, pytorch) – будет ли готовая модель, какое решение уже существует?

1) Сбор и обработка данных: первый этап - сбор и очистка данных для обучения модели.

2) Разделение на тренировочный и тестовый наборы

3) Выбор модели

4) Обучение модели: на этом этапе модель получает данные из тренировочного набора и настраивает свои параметры, чтобы оптимизировать процесс обучения на этих данных.

5) Оценка качества модели на тестовой выборке.

6) Тонкая настройка и улучшение модели: лучшие результаты можно достичь с помощью тонкой настройки модели, что включает в себя различные техники, например, регуляризацию, оптимизацию параметров модели, а также выбор новых алгоритмов для улучшения ее производительности.

4. Сравнительный анализ метрики качества для всех датасетов и выявление “некачественных” датасетов.(Python, pandas, numpy, scipy, sklearn, pytorch)

5. Разработка алгоритма приведение данных к качественной форме/виду. (Python, pandas, numpy, scipy, sklearn, pytorch)

6. Тестирование алгоритма. Оценка статзначимости результата работы модели на обработанных данных по сравнеию с исходными

7. Разработка пайплайна для оценки качества датасета и реализации алгоритма его редакции (Python, pandas, numpy, scipy, sklearn, pytorch, pytorch lightining, piplines)

8. Разработка инфраструктуры сервиса?

Драфт метрик и критериев успеха: статзначимый результат работы модели на обработанных данных по сравнеию с исходными.

План возможных рисков и как их избежать:

- полученная метрика не оценивает данные адекватно (например, показывает оинаковые значения для всех датасетов). Решение: возврат к исходной точке анализа и поиск новых метрик

- части проекта не будут собраны воедино. Решение: work hard

- возможно для каких-то датасетов алгоритм редакции данных не даст результата. Решение: возможно данные сильно зашумлены и алгоритм не подходит, возможно данных недостаточно и нужно дополнить данные.

- не будет понимания о том как были собраны данные и в связи с этим препроцессинг может быть не адекватным. Решение: пойти к заказчику и спросить.

* Шесть ключевых показателей качества данных для автономного мониторинга включают в себя точность, полноту, согласованность, своевременность, уникальность и достоверность.



### 1. Точность

Точность измеряет количество и типы ошибок в наборе данных. Отслеживание этой метрики отвечает на основной вопрос: находится ли поведение данных в ожидаемых пределах?

В наборе данных могут присутствовать различные типы неточностей, в том числе:

* **Аномальное значение:** находится ли значение в поле в допустимых пределах.
* **Аномальный микрообъем:** находится ли объем записей в микросегменте данных в допустимых пределах.
* **Аномальная строка:** соответствует ли значение в поле параметрам в предварительно определенном списке.
* **Аномальные отношения между столбцами:** точны ли отношения между несколькими столбцами в записи
* **Аномальное распределение:** является ли распределение данных в каждом микросегменте точным.

Для измерения точности данные можно сравнить с эталонным набором данных. Например, введенные клиентом адреса и почтовые индексы можно сравнить с реестром адресов USPS. Его также можно проверить логически, чтобы увидеть, попадает ли он в принятые границы и имеет ли он смысл.

Точность обычно измеряется как процент ошибок к общему количеству записей. Более низкий процент указывает на более высокое качество данных.

### 2. Полнота

Важно, чтобы все критические поля в записи были полностью заполнены. Например, запись клиента, в которой отсутствует почтовый индекс, является неполной.

Метрика полноты измеряет количество записей с неполными данными. Он отслеживается путем выявления записей с пустыми полями и обычно выражается в процентах от общего количества записей.

Обратите внимание, что некоторые поля могут быть пустыми — например, поле, которое запрашивает необязательный отзыв клиента. Значение NULL в поле этого типа не должно включаться в показатель полноты.

### 3. Консистенция

Согласованность измеряет, как отдельные точки данных, извлеченные из двух или более наборов данных, синхронизируются друг с другом. Если две точки данных конфликтуют, это указывает на то, что одна из обеих записей неточна.

Например, вы можете объединить данные о клиентах из двух разных источников. При объединении данных для данного клиента вы хотите, чтобы все поля были согласованы — клиент должен иметь тот же адрес улицы, почтовый индекс и номер телефона. Если два источника имеют разные номера телефонов одного и того же клиента, данные несовместимы.

Несоответствие может иметь ряд причин, в том числе:

* Неверно введены данные в одном или нескольких источниках
* Данные, введенные по-разному в одном или нескольких источниках (например, «Cir» или «Circle» в почтовом адресе)
* Данные получены в разное время (указание на то, что более новые данные могут отражать изменения по сравнению с более ранними данными)
* Различная структура/схема между источниками данных не полностью соответствует

### 4. Своевременность

Своевременность, иногда называемая валютой, — это показатель, который измеряет возраст данных в базе данных. Более свежие данные, вероятно, будут более точными и актуальными, поскольку информация, особенно информация о клиентах, может меняться со временем. Кроме того, существует значительный риск увеличения количества ошибок при перемещении старых данных по конвейеру, поскольку все промежуточные репозитории данных заполняются результатами из устаревших данных.

Своевременность отслеживается с помощью временных меток в источниках данных. Установите базовую разницу между текущей датой, а затем повторно проверьте, обновите или заархивируйте старые данные.

### 5. Уникальность

Метрика уникальности отслеживает повторяющиеся данные. Вы не хотите подсчитывать какие-либо данные дважды, так как это может чрезмерно взвесить любые результаты. Важно определить дубликаты и либо объединить их, либо удалить дубликаты.

### 6. Срок действия

Валидность измеряет, насколько хорошо данные соответствуют стандартам. То есть данные, введенные в данное поле, должны быть надлежащего типа; если введен неправильный формат данных, эти данные могут оказаться непригодными для использования.

Важно убедиться, что каждый фрагмент данных соответствует правильному типу и требуемому формату данных. Например, двузначный буквенный код штата не может содержать три буквы или какие-либо цифры. Поля даты должны соответствовать установленному формату, например, число/месяц/год, а не какому-либо другому порядку. Неверные данные должны быть идентифицированы и исправлены.