Департамент образования и науки города Москвы
Государственное автономное образовательное учреждение
высшего образования города Москвы
«Московский городской педагогический университет»
Институт цифрового образования
Департамент информатики управления и технологий

Нургалеева Гузель Рустэмовна БД-241м

Инструменты хранения и анализа больших данных

Практическая работа 3.1. Анализ и визуализация больших данных. Машинное обучение на больших данных с использованием Apache Spark MLlib

Вариант задания: 18

Направление подготовки/специальность

38.04.05 - Бизнес-информатика

Бизнес-аналитика и большие данные

(очная форма обучения)

Москва

2025

Задание 18.

1) Опишите шаги, предпринятые в разделе 3 для первичной оценки данных (`user_activity_workout_summarize`). Почему важно получить эти общие цифры перед углубленным анализом?

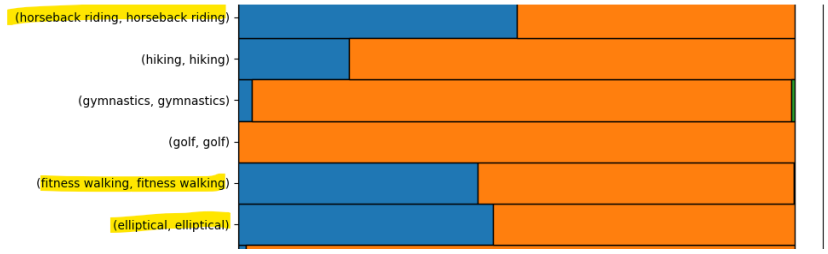
Проверка на пропущены значения и аномалии важна т.к.
- могут плохо отразиться на моделях машинного обучения
- пропущенные значения могут быть индикатором неверно собранного датасета (мог произойти технический сбой при выгрузке данных)
- большое количество пропущенных данных может послужить поводом для пересмотра набора параметров в датасете
- аномалии стоит собрать в отдельный датасет и попробовать найти закономерности их появлений

Также исходя из того сколько всего записей в наборе, сколько пользователей и вариантов показателей других параметров в наборе данных можно определить
- достаточность записей
- необходимость (возможность) разбить на поднаборы для анализа
- необходимость дополнения набора новыми параметрами или записями
- определить примерное время на проведение анализа

2) Проанализируйте stacked bar chart гендерного распределения по видам спорта (раздел 6). Назовите 2-3 вида спорта с наиболее сбалансированным гендерным составом.

Данный график дает возможность быстро понять как распределяются данные по видам спорта между полами.
В большинстве видов спорта бОльшая часть данных получена от пользователей мужского пола.
Также есть вид спорта «Танцы», для которого пол не определен. Что в данном случае может считаться аномалией (стоит проверить не было ли сбоя в базе данных, в передаче данных с устройств и т.д.).

Из представленного списка следующие виды спорта являются наиболее сбалансированными с точки зрения гендерного состава:
- horseback riding
- fitness walking
- elliptical



3) Интерпретируйте 3D-график для 'orienteering' (раздел 8). Что необычного в траектории этого вида спорта по сравнению, например, с 'running'?

В ориентировании
- больше разброс данных в 95 процентили
- среднее значение находится в пределах 75 процентили
- ориентирование проходит быстрее, чем бег. Т.к. 10 сек значение приходится на 75 процентиль ориентирования

4) **Напишите код PySpark, чтобы создать новый датафрейм, содержащий только userId тех пользователей, которые занимались 'running' И 'cycling'.**

(предложенные виды спорта исправила на ‘run’, ‘indoor cycling’ согласно данным датасета)

```
# Получаем список всех unique users, кто занимается бегом
runners_df = df.filter(df['sport'] == 'run').select('userId').distinct()

# Получаем список всех unique users, кто занимается сайклингом
cyclists_df = df.filter(df['sport'] == 'indoor cycling').select('userId').distinct()

# Выполняем внутреннее соединение между двумя наборами данных, чтобы выбрать только тех,
# кто делает и то, и другое
result_df = runners_df.join(cyclists_df, on='userId', how='inner')
```

```
# Сохраняем результат в новую таблицу
final result = result df.distinct()
# Показываем результат
final_result.show()
```

результат выполнения кода

```
cyclists_df = df.filter(df['sport'] == 'indoor cycling').select('userId').distinct()

# Выполняем внутреннее соединение между двумя наборами данных, чтобы выбрать только те
# кто делает и то, и другое
result_df = runners_df.join(cyclists_df, on='userId', how='inner')

# Сохраняем результат в новую таблицу
final_result = result_df.distinct()
# Показываем результат
final_result.show()
```

```
+--------+
|  userId|
+--------+
| 8467445|
| 5325166|
|  897592|
| 1655221|
|  982359|
|14066832|
|  334217|
| 5255745|
| 3559941|
| 7470676|
| 6584414|
| 2868369|
| 5964610|
|  260784|
|  993718|
| 2486861|
| 1957863|
| 2675116|
|  517351|
| 3700284|
+--------+
only showing top 20 rows
```

5) Предложите идею ML-модели в Spark MLlib для прогнозирования оттока пользователей (churn prediction). Какие признаки, отражающие активность пользователя (частота тренировок, разнообразие видов спорта, динамика продолжительности), могли бы быть ключевыми? Какая метрика оценки модели была бы важна?

Нужно определить для каждого вида спорта:
- частоту тренировок
- LT(live time) по виду спорта (т.е. через какое время после первой тренировки посетители перестают посещать занятия, т.е. происходит отток)
- время, которое прошло с момента последней тренировки до дня проведения анализа
- гендер (возможно, частота посещения и LT зависят от гендера)
- успеваемость (прогресс, регресс). Возможно, клиенты уходят, когда прогресса больше нет или если начинается регресс в показателях
- если есть информация о том, кто из посетителей ушел, т.е. фактический churn , то использовать этот признак. И используя его определить остальные важные признаки при помощи корреляционного анализа

Важной метрикой оценки модели может быть F1 score
F1 score — балансировка Precision и Recall, представляет собой среднее гармоническое между ними

Либо **Recall** (т.к. отражает насколько хорошо модель определяет случаи оттока), важно для того, чтобы вовремя его предотвратить.