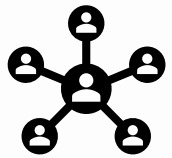
Deloitte.



Entity Duplication Detection

Analytathon 2

26th April - 2nd March

Introductions

Deloitte.

Sarah Crawford

Manager

Data Scientist specialising in **Machine Learning** & **Geospatial Products**

Matthew Orr

Consultant

Data Scientist specialising in **Geospatial** & **Simulation Engineering**

QUB MSc Data Analytics Graduate

Overview of the Challenge

Deloitte.

Definition of the issue to be solved

Background

A leading organisation is facing a critical **data quality challenge** within its customer management system (CMS). **Duplicate customer records** have proliferated, creating a tangled web of inconsistencies that hinder operational efficiency, impact strategic decision-making, and threaten the organization's data-driven aspirations.

Challenge

Develop a data-driven solution to **identify & manage potential duplicate customer records** within the CMS. You should report on the number of duplicated records and provide recommendations to prevent duplication at source. Your solution should leverage data science, harness analytical techniques, and demonstrate a deep understanding of data quality principles.

The Problem’s Impact

Where the problem occurs in the real-world

Duplicate customer data can

- … lead to inaccurate sales forecasts and an inability to effectively target and nurture customer relationships.
- … cripple operational efficiency by wasting employee time on reconciliation and hindering process automation.
- … undermine customer experience, leading to frustration from inconsistent interactions and a negative perception of the brand's competence.
- … raise significant compliance concerns, increasing the risk of data breaches, privacy violations (like GDPR), and inaccurate reporting that could lead to penalties.

The Dataset

Deloitte.

Information on the record dataset

Information Contained

- Personal, contact, and location information on customers
    - **Not real/live data**
- Around 5,000 total records in the dataset
- An unknown number of duplicates
    - No ground truth / labels

Data Inconsistencies

Be mindful of data inconsistencies that exist in the database

- *How should data cleansing be handled?*

Columns
prefix
first_name
middle_name
last_name
date_of_birth
email_address
phone_number
house_no
primary_street
town
postcode
county

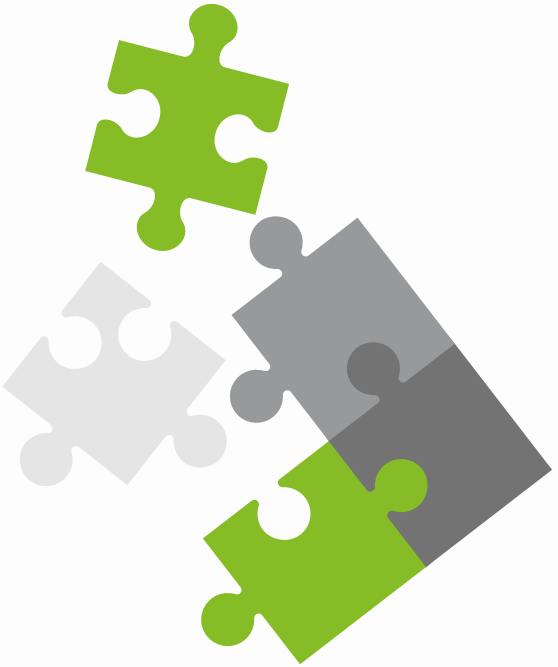
Dataset Issues

Deloitte.

Data quality issues within the challenge's dataset

Potential issues with the dataset

- Missing data throughout the columns
- Email address inconsistencies
    - People may personally have more than one
- Phone number formats
- Data entry error
    - Typos / character entry
- Names aren't always a strong duplicate identifier
    - People with the same name
    - People that interchange their first name



Potential Solutions

Possible avenues to achieve results

Deloitte.

Possible Approaches

- Analytical rule-based approach?
- Harness pattern matching algorithms?
- Unsupervised ML model?

Considerations

- Could you score the detected matches?
- Are some variables more important?
- Are there ways to discard obvious non-matches?

Key Takeaways

- There is no correct answer – only strong solutions
- Experiment with different approaches
- Encourage logical & analytical thinking

Deloitte.

Good luck!

Thank you for listening, any questions?