GROUP 8

# Analytathon 2

# Entity Duplication Detection

Bhagyaprasad Vastrad

# Overview

We developed a **data deduplication model** to identify and remove duplicate records.

The approach leverages **data preprocessing, fuzzy matching, and thresholding** to detect duplicates.

Our goal is to **enhance data quality and improve business efficiency**.

# Exploratory Data Analysis

## Data Preprocessing:

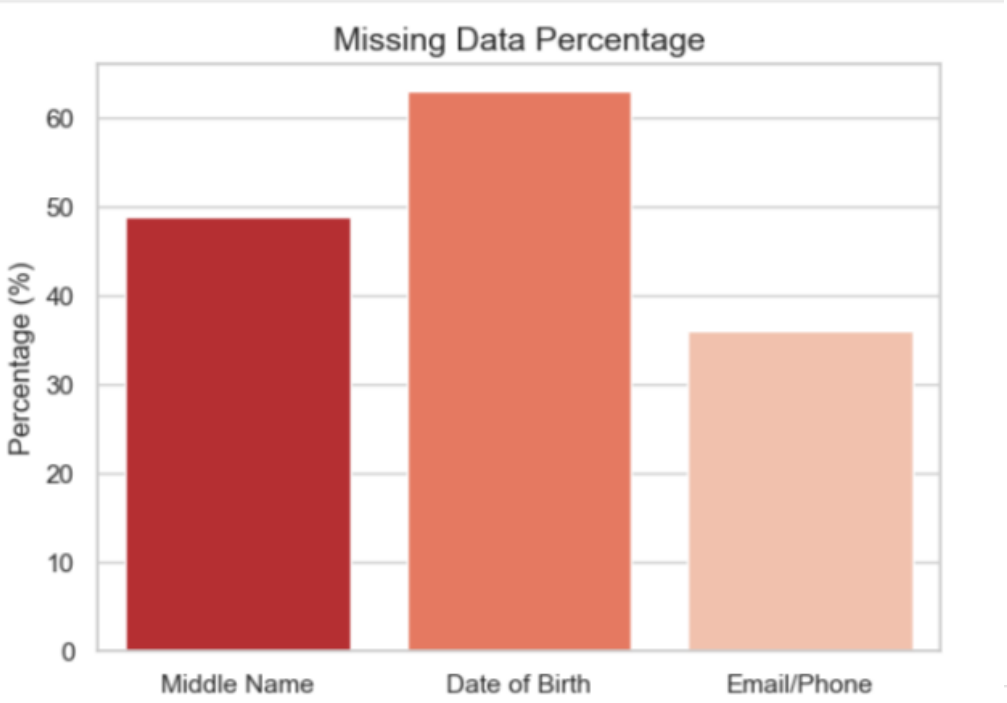
- Standardizing date_of_birth.
- Normalizing email_address (lowercase, removing spaces).
- Standardizing phone_number (removing special characters, ensuring uniformity).

## Duplicate Detection:

- **Fuzzy Matching:** Uses fuzzywuzzy to compare string similarities.
- **Pairwise comparison** to find close matches.
- **Similarity thresholding** to detect duplicates with confidence.

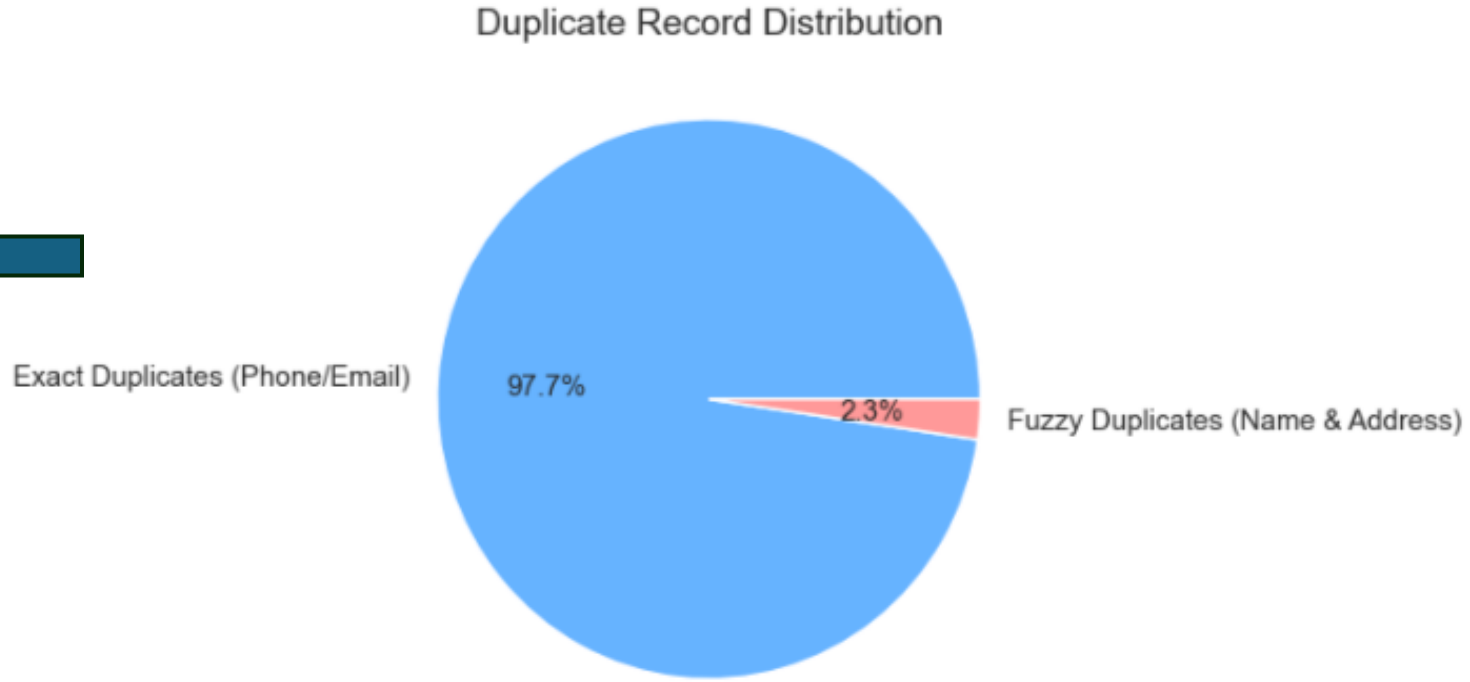
# Duplicate Detection Summary

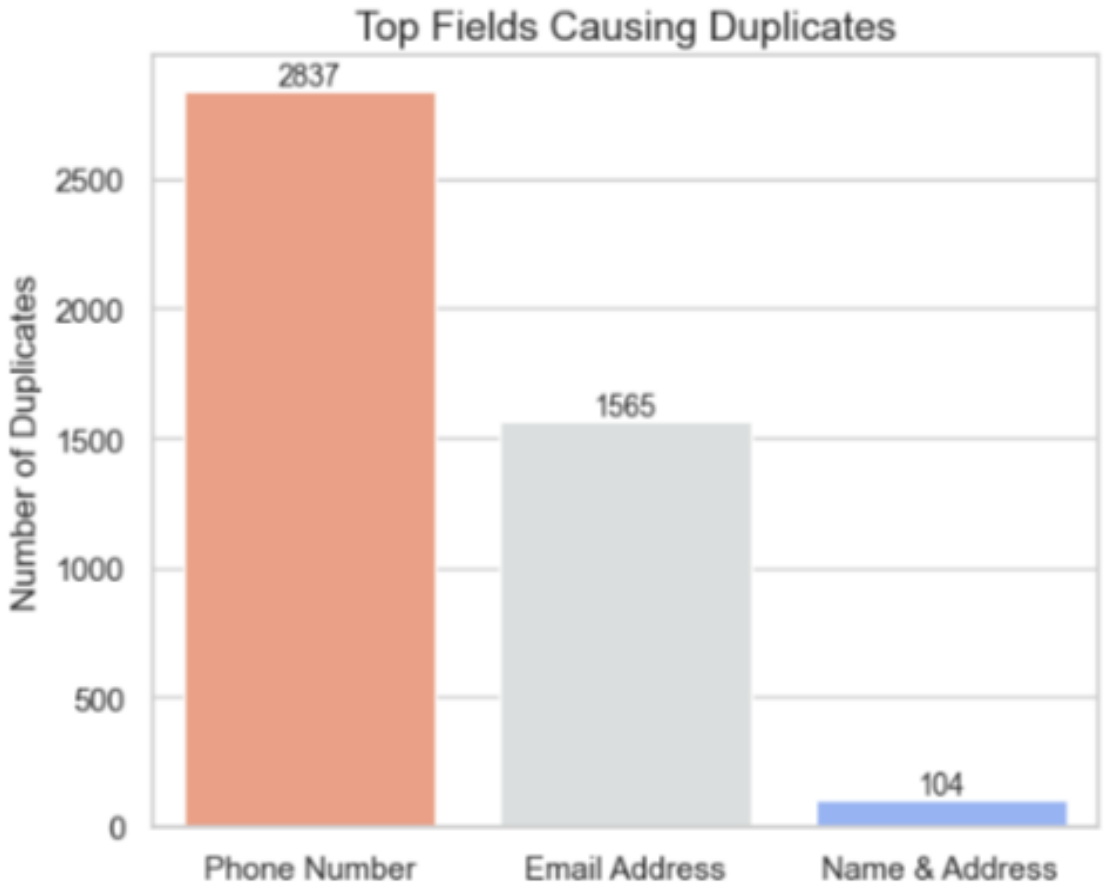
| Category | Count |
|---|---|
| Total Records Processed | 5000 |
| Missing Middle Name | 49% missing |
| Missing Date of Birth | 63% missing |
| Missing Email/Phone | ~36% missing |
| Exact Duplicates (Phone-based) | 2837 |
| Exact Duplicates (Email-based) | 1565 |
| Fuzzy Duplicates (Name & Address) | 104 |
| Total Duplicates Identified | 3512 |




Date of Birth has 63% missing values, making identity matching difficult.
Email/Phone is missing in 36% of records, reducing duplicate detection accuracy.

Most duplicates (3512 records) are exact matches based on phone/email.
Fuzzy duplicates (104 records) suggest minor differences in name or address.






- Phone Number is the most common duplication source, causing 2837 duplicate records.
- Email Address is the second biggest factor, contributing to 1565 duplicates.

# Fuzzy Matching

Fuzzy matching is a technique used to identify similar but not identical records within a dataset.

It helps detect duplicate entries even when there are minor differences in names, addresses, or contact details.

This method is useful for handling inconsistencies caused by typos, varying formats, and incomplete data.

# Why  Is Fuzzy Matching Necessary?

## Challenges

**Varied Spellings:** Accommodate different spellings of the same name or term.

**Formatting Inconsistencies:** Handle variations in address or phone number formats.

**Data Gaps:** Match records despite missing information.

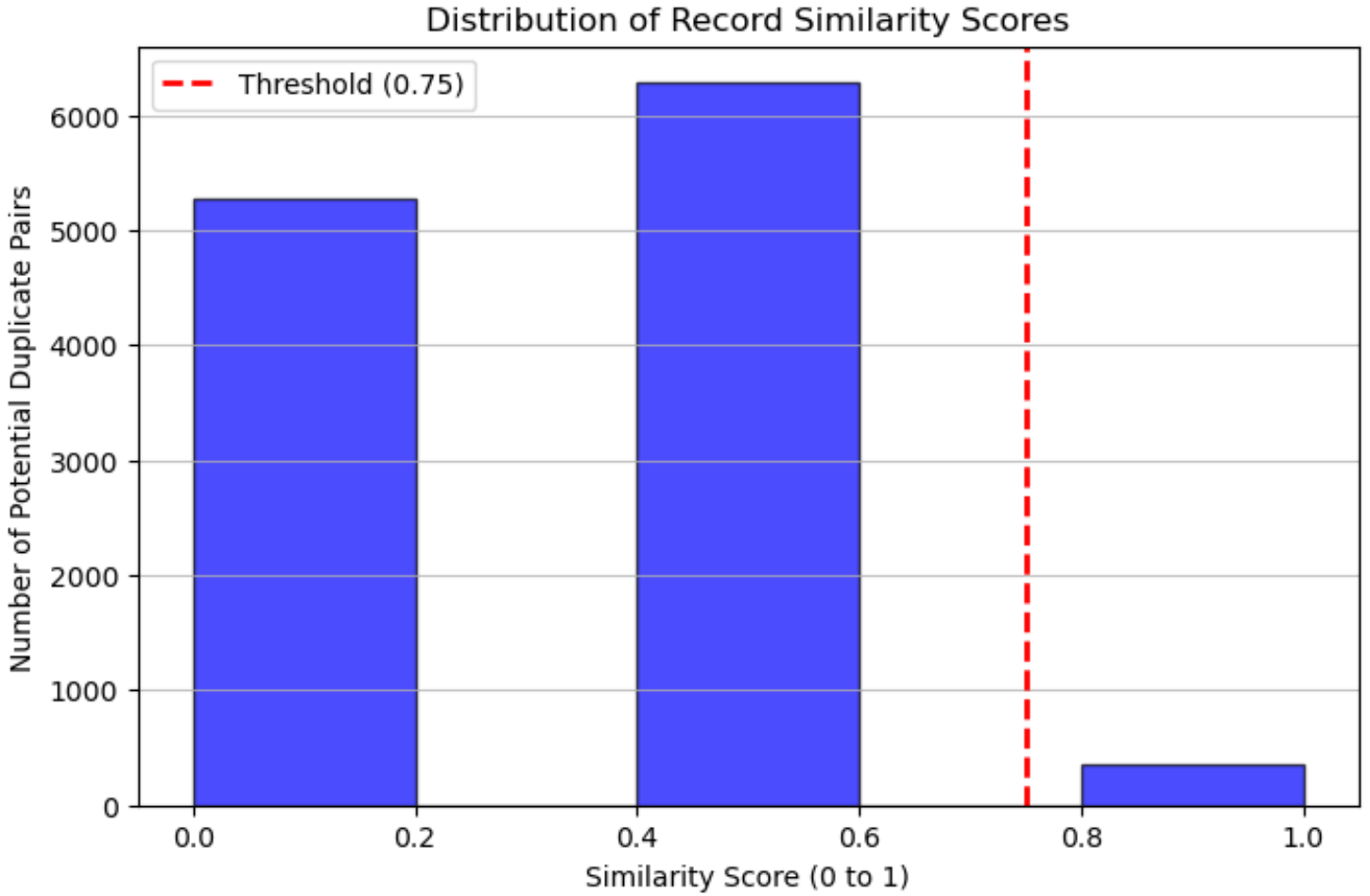
## Consequences

- **Inefficient Operations:** Redundant efforts and wasted resources.
- **Skewed Insights:** Misleading analysis due to inaccurate data.
- **Compliance Issues:** Legal and regulatory repercussions from data errors.

# Steps in our approach

- **Removing Exact Duplicates:** Identifying records with the same email or phone number and eliminating them.
- **Generating Potential Matches:** Using **Sorted Neighborhood Indexing** on first names to efficiently pair similar records.
- **Calculating Similarity Scores:** Assigning a weighted score based on:
  - First Name Similarity (40%)
  - Last Name Similarity (40%)
  - Primary Street Similarity (20%)
- **Filtering Duplicates:** Records with a **similarity score above 0.75** are considered probable duplicates.

# Performance and Statistics - Similarity Score




Similarity Threshold: 0.75

# Similarity Threshold

- The similarity score measures how closely two records match, ranging from **0 (completely different)** to **1 (identical)**.

- We set a threshold of **0.75**, meaning only records with an **75% or higher similarity** are considered potential duplicates.

- If the threshold is **too low** (e.g., 0.6), we mistakenly label unique records as duplicates, leading to errors.

- If the threshold is **too high** (e.g., 0.95), we fail to catch actual duplicates, keeping redundant records in our system.

| | Record 1 | Record 2 |
|---|---|---|
| prefix | mrs | miss |
| first_name | deborah | deborah |
| middle_name | | shannon |
| last_name | chamberlain | watson |
| DOB | | |
| email | | |
| phone_number | | |
| House_number | 187 | 116 |
| Primary_street | gilbert rue | ross lights |
| town | | ardgarvan |
| county | derry | tyrone |
| Post code | BT51 3ES | BT78 9KS |
| | | |

**Low Similarity** (0.4):
Likely NOT a Duplicate

| | Record 1 | Record 2 |
|---|---|---|
| prefix | mrs | mrs |
| first_name | eielen | eileen |
| middle_name | | |
| last_name | pope | pope |
| DOB | | |
| email | | |
| phone_number | | |
| House_number | | |
| Primary_street | roy ridge | roy ridge |
| town | drumnakilly | drumnakilly |
| county | fermanagh | fermanagh |
| Post code | BT92 5WA | BT92 5WA |
| | | |

**High Similarity** (1.0): Likely a Duplicate

# Solutions

What can be done to avoid duplicate entries?

- **Implement Standardized Data Entry**: Ensure uniform formatting for names, phone numbers, and addresses.
- **Introduce Real-time Duplicate Detection:** Validate customer data as it is entered.
- **Use Unique Identifiers:** Assign customer IDs instead of relying solely on names.

# Key Findings

- Our analysis flagged 3,616 duplicates out of 5,186 which means we detected that over 75% of records were duplicates.
- The majority of these were exact matches based on phone numbers and email addresses (3,512) and then we captured a further 104 duplicates using fuzzy  matching.
- A significant amount of missing data in the date of birth and email/phone fields complicated analysis.
- We used a combination of data preprocessing, rule-based exact matching and weighted fuzzy matching which proved effective in our detection.