# **Persona Matching – Modelling Framework**

## **Hybrid Modelling Approach:**

The persona matching engine targeted 9 types of problem, which falls in the fuzzy matching realm.

1. Textual Similarity (typos, etc.)
2. Nickname (Mike/Michael)
3. Missing spaces/hyphens (Mary Ellen/MaryEllen/Marry-Ellen)
4. Initials (JE Smith/James Earl Smith)
5. Out of order components (Diaz Carlos Alfonzo/Carlos Alfonzo Diaz)
6. Name Swap (Jesse Jones/Jones Jesse) – **Will be further improved different** Name Split (First: Dick, Last: Van Dyke/First: Dick Van, Last: Dyke)
7. Truncated Name (Charles Livingston/Charles Living)
8. Missing Name (John Albert Lewis Lewis/John Lewis)
9. Maiden name addition (or any other additional last name)
10. Phonetics Similarity will be taken in the future – **Will be considered in the future**
11. **Un-Supervised Learning Approach -** Steps undertaken to Improve Model Accuracy we used following steps

Understood the percentage of data distribution scenario:

1. Director\_name: 100%
2. DOB: 65%
3. Visa\_no: 37%
4. Mobile\_1: 39%
5. Mobile\_2: 22%
6. National\_id\_number: 0.01%

***Vector length will be a combination of director\_name + {‘feature set’}***

1. **Scenario 1:** director\_name + {‘mobile\_1’ , ‘mobile\_2’}
2. **Scenario 2:** director\_name + {‘passport\_number’, ‘Age’}
3. **Scenario 3:** director\_name + {‘Age’, ‘passport’, ‘visa’, ‘mobile\_1’, ‘mobile\_2’}
4. **Scenario 4:** only director\_name. These are those records that have no data for the specific director. In other words, the entire feature set is null.

**(B) Supervised Model/ Spam Detection Model (Future Scope):** Extracted various business suggestions and systematically inputted the same to attain better match rate.

Feedback from Unsupervised approach model would be leveraged in the supervised model. Thus, this approached of persona matching engine will further optimize the 10 types of above-mentioned problems.

Here, we are planning to label the data and convert it into a classification problem, which will require human cognition. Therefore, it will allow us to apply GBM, XGBoost and Random Forest.

For example, if there is a match e.g. Jennifer Williams/Jenny Williams; labeled as 1, and cases that are “close” match eg. Don Anderson/Donan Anderson, but are different; labeled as 0.

*Note: There are various enhancements which require lot of R&D work:*

* *Apply Machine Learning to automatically identify the Spam.*
* *Phonetics Similarity will be taken in the future*

## **Modeling Solution Steps:**

**Step 1:** Understood the percentage of data distribution scenario:

1. Director\_name: 100%
2. DOB: 65%
3. Visa\_no: 37%
4. Mobile\_1: 39%
5. Mobile\_2: 22%
6. E\_mail: 22%
7. National\_id\_number: 0.01%

**Step 2:** Also undertaken lot of preprocessing steps to enhance the quality of data:

1. The dob has numerous junk and incorrect human feed data. We developed a logic that will consider only ages >=15 and convert date to age to resolve this issue.
2. We handled all the null values across the feature set as mentioned above.
3. Moreover, there are numerous places where the user has added short forms, which are also being handled. For example:

SANA ABDUL XYZ(N/R), SANA ABDUL(UBO) XYZ, SANA ABDUL XYZ(Country name), SANA(NR) ABDUL XYZ etc.

1. Also, we handled the titles given to each individual, which is also decreasing the quality of data. Such as MR, MRS, HH, HE
2. All the special characters such as ‘[:./,#%?\=@”\*]’, and others have also been handled.
3. We also handled digits in the name, such as ‘01043613015 abc’ , ‘abc 234 ertt’.

**Step 3:** Four case scenarios developed to cover all the text matching problems mentioned above.

***Vector length will be a combination of director\_name + {‘feature set’}***

1. **Scenario 1:** director\_name + {‘mobile\_1’ , ‘mobile\_2’}
2. **Scenario 2:** director\_name + {‘passport\_number’, ‘Age’}
3. **Scenario 3:** director\_name + {‘Age’, ‘passport’, ‘visa’, ‘mobile\_1’, ‘mobile\_2’}
4. **Scenario 4:** only director\_name. These are those records that have no data for the specific director. In other words, the entire feature set is null.

## **Modelling Output Validation by Business:**

We had undertaken a sample across

|  |  |  |
| --- | --- | --- |
| **Sample Size - Analyzed** | **Total Population** | **Sample %** |
| 7284 | 56038 | 13.00% |

Stats based on the above sample are given below:

|  |  |
| --- | --- |
| **Model Validation Metrics** | **% Improvement** |
| Overall False Positive | 99.92% |
| Existing Dedupe of Director Name | 94% |
| Further Enhancement required to remove dedupe | 6% |

False Positive Distribution: Total is 6 – (UAE – 2 and Egypt – 4 )