Entity Resolution

A Machine Learning use case

Agenda

- Problem Statement
- Data Preparation
- Feature creation - Similarity Function
- Feature Set
- Combining Feature set across record – Data Model
- Outcome - Testing the Model
- Summary – Data Model Life Cycle

Problem Statement

Entity Resolution

Resolving different entities into a single customer view using Machine Learning.

- Due to multiple systems and multiple booking channels in place there could be duplication of same entities, for e.g.:

ID	NAME	Address	Birth_Date	Email
1	Niamh	Perth, AU	7-Aug-1968	footloose.xcape@gmail.com
2	Niamh Mary	Perth, AU	7-Aug-1968	-
		…		
140	Wai Ping	Brisbane, AU	19-Nov-1956	ccsoong@gmail.com
141	Wai	Newcastle, AU	1-Jan-1900	ccsoong@gmail.com

Data Preparation

Training Dataset

A sample dataset is manually labelled to train and tune the Machine Learning model. Here, we picked up a set of 1000 records that contains multiple duplicate entities.

Below are the fields chosen for this study:

ID, PREFERRED_NAME, NAME, SURNAME, SEX, BIRTH_DATE, HOME_COUNTRY, EMAIL, ACCOUNT_STATUS, ACTIVATION_DATE, HOME_ADDR, HOME_PCODE, HOME_SUBURB, HOME_STATE, SCI_ID

PII fields were chosen for analysis

Sample Labelled Data

ID	NAME	Address	Birth_Date	Email	SCI_ID
1	Niamh	Perth, AU	7-Aug-1968	footloose.xcape@gmail.com	41
2	Niamh Mary	Perth, AU	7-Aug-1968	-	41

140	Wai Ping	Brisbane, AU	19-Nov-1956	ccsoong@gmail.com	161
141	Wai	Newcastle, AU	1-Jan-1900	ccsoong@gmail.com	161

Paired sets are manually labelled

Feature Creation - Similarity Function

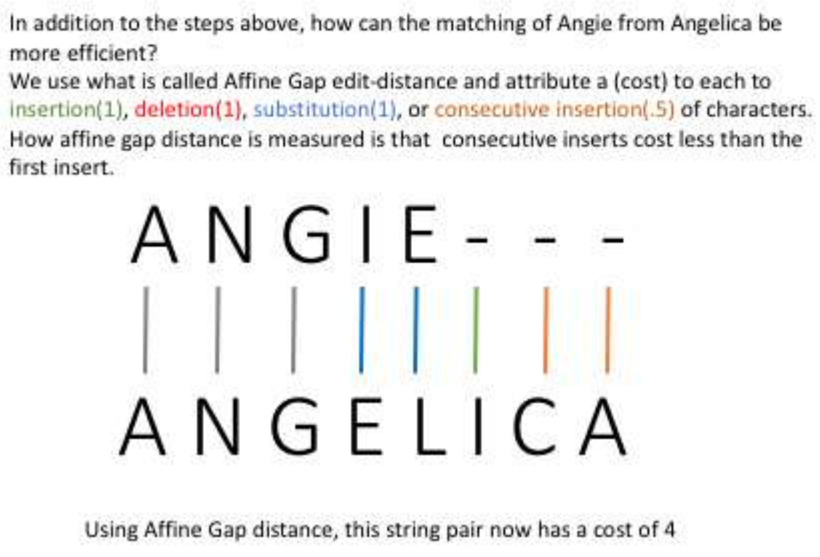
Similarity Function

It is desirable to learn similarity functions from training data to capture the correct notion of distance for a particular task in a given domain.

Different Algorithms adopted to compute the similarity across the records are:

levenshtein, Jaro, Jarowinkler, Damerau_levenshtein, qgram, cosine, smith_waterman, lcs

Illustration - Affine Gap



Feature Set

		Pref_name_lev	Pref_name_jaro	Pref_name_lcs	Pref_name_dam	...	HOME_ADDR_lcs	Match
ID_1	**ID_2**							
1	**3**	0.125	0.383333	0.383333	0.125	...	0.27027	1
1	**4**	0.2	0.522222	0.522222	0.2	...	0.285714	0
1	**5**	0.142857	0.447619	0.447619	0.142857	...	0.341463	0
1	**6**	0.142857	0.447619	0.447619	0.142857	...	0.318182	0

ID Pair

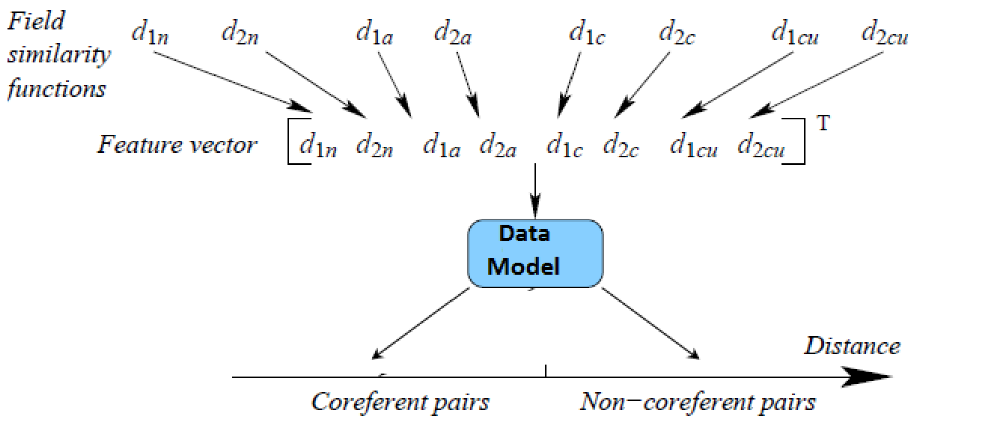
Feature

Label

Combining Similarity Function – Data Model

Data models treat individual field similarities as features and train a classifier to distinguish between coreferent and non-coreferent records, using the confidence of the classifier's prediction as the similarity estimate.

Name	*Address*	*City*	*Cuisine*
Fenix	8358 Sunset Blvd. West	Hollywood	American
Fenix at the Argyle	8358 Sunset Blvd.	W. Hollywood	French (new)



Outcome – Testing the Model

Data Model

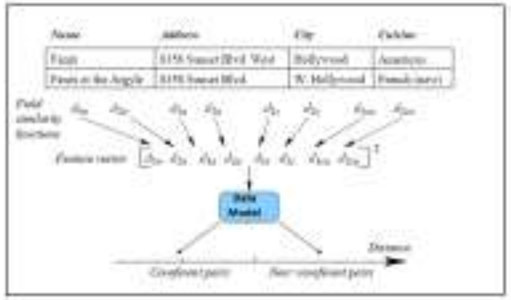
	Actual (No Match)	Actual (Match)	
Predicted (No Match)	385827	69	922
Predicted (Match)	711	151	34
	201	755	

The Model was tested with different set of records and an accuracy of 99.8% was recorded.

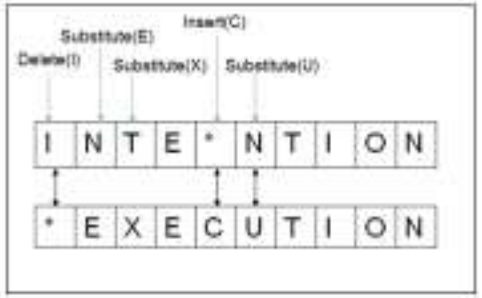
ER Data Model

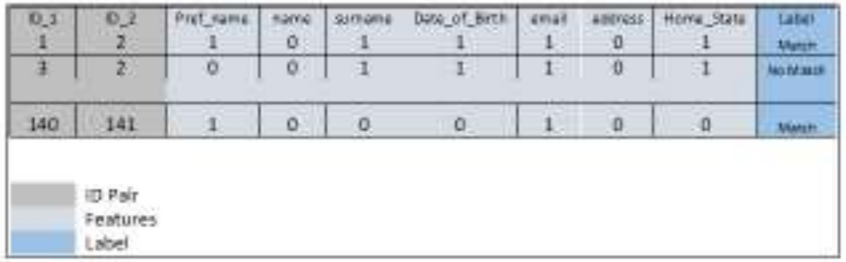
Summary – Data Model Lifecycle

Axiom					
ID	Name	Address	Email	Bith_date	SCI_ID
1	Niamh	Perth, AU	footloose.xcape@gmail.com	1968-08-07	41
2	Niamh Mary	Perth, AU	-	-	41
...					
140	Wai Ping	Brisbane, AU	ccsoong@gmail.com	1956-11-19	161
141	Wai	Newcastle, AU	ccsoong@gmail.com	1900-01-01	161



Data Modeling Life cycle





ID_1	ID_2	Pref_name	name	surname	Date_of_Birth	email	address	Home_State	Label
1	2	1	0	1	1	1	0	1	Match
3	2	0	0	1	1	1	0	1	No Match
140	141	1	0	0	0	1	0	0	Match

ID Pair
Features
Label

THANK YOU !