Evaluation of Data Matching in the Immigration Flow Dataset

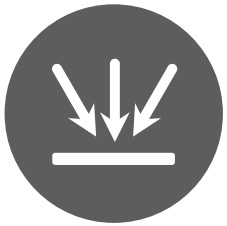
Hongwei Zhang, Jiashen You, Andrea Bonilla, Katie Shanahan, Andrew Leach

Agenda

- Introduction
- Data
- Approach
- Business Rules
- Probabilistic Models
- Conclusion


Introduction
Data
Approach
Business Rules
Probabilistic Models
Conclusion

Introduction



OIS **combines data from ICE, CBP, USCIS, and DOJ/EOIR** to report on U.S. immigration system

Data are all **event-based**, but an individual can experience **many events across components**

Need a **complete picture** of how people move **through the immigration system**

Identifiers and data quality vary across sources and datasets

OIS combined 13 datasets to create an 38 million record **flow dataset** and needs to match by person


Introduction
Data
Approach
Business Rules
Probabilistic Models
Conclusion

Data Sources in the Flow Dataset

U.S. Citizenship and Immigration Services (USCIS)

- Affirmative Asylum
- Defensive Asylum
- Deferred Action for Childhood Arrivals (DACA)
- Legal Permanent Residents (LPR)

Immigration and Customs Enforcement (ICE)
Enforcement and Removal Operations (ERO)

- Charging Documents Issued/Notice to Appear (NTA)
- ERO Administrative Arrests
- ERO Intakes and Releases (Detention)

ICE Homeland Security Investigations (HSI)

- HSI Administrative Arrests

Customs and Border Protection (CBP)
Office of Field Operations (OFO)

- OFO Inadmissibles

CBP U.S. Border Patrol (USBP)

- USBP Apprehensions

DHS

- Removals and Returns

DOJ Executive Office for Immigration Review (EOIR)

- EOIR

The flow dataset is sourced from 13 datasets
obtained from seven components across DHS and DOJ

Numeric Identifiers

Identifier	Description	Level	Assessment
Alien File Number (anumber)	Unique identifier used across DHS immigration components and EOIR	Person	▪ Very strong indicator ▪ Some individuals have multiple anumbers ▪ Missing or invalid (000000000) anumbers
EID civ ID	Identifier generated by the Enforcement Integrated Database	Encounter	▪ Used only by ICE and CBP ▪ Encounter level; cannot be used to match people across encounters ▪ Includes some missing values and errors
Ident FIN	Automated Biometric Identification System (ADIS) ID, generated from fingerprints	Person	▪ Very strong identifier ▪ OIS does not receive Ident FINS for USCIS or EOIR data

There are three available numeric identifiers in the datasets, though none are individually sufficient for matching

Biographic Identifiers

Identifiers and Common Errors

Name

- Different names included (first vs. first and middle)
- Multiple first, middle, or last names
- Different spellings, nicknames, typos
- Hyphens vs. spaces
- Common suffixes and prefixes (“de la”) may or may not be included
- Accented letters

Birth Date

- Month and day swapped
- January 1 is a common filler date
- Typos and errors

Gender

- Sometimes incorrect
- Sometimes unknown or missing in one dataset and M or F in another

Citizenship

- Sometimes incorrect
- Sometimes missing
- May vary between datasets; individual may have multiple citizenships, change citizenship, or lie

Datasets also include four biographic identifiers,
but errors associated with each pose challenges for matching


Introduction
Data
Approach
Business Rules
Probabilistic Models
Conclusion

Method Selection

Rules-Based Methods

Sorted Neighborhood Method
Blocking Technique
Merging

Probabilistic Methods

Logistic Regression
Naïve Bayes
Support Vector Machine (SVM)
Decision Tree

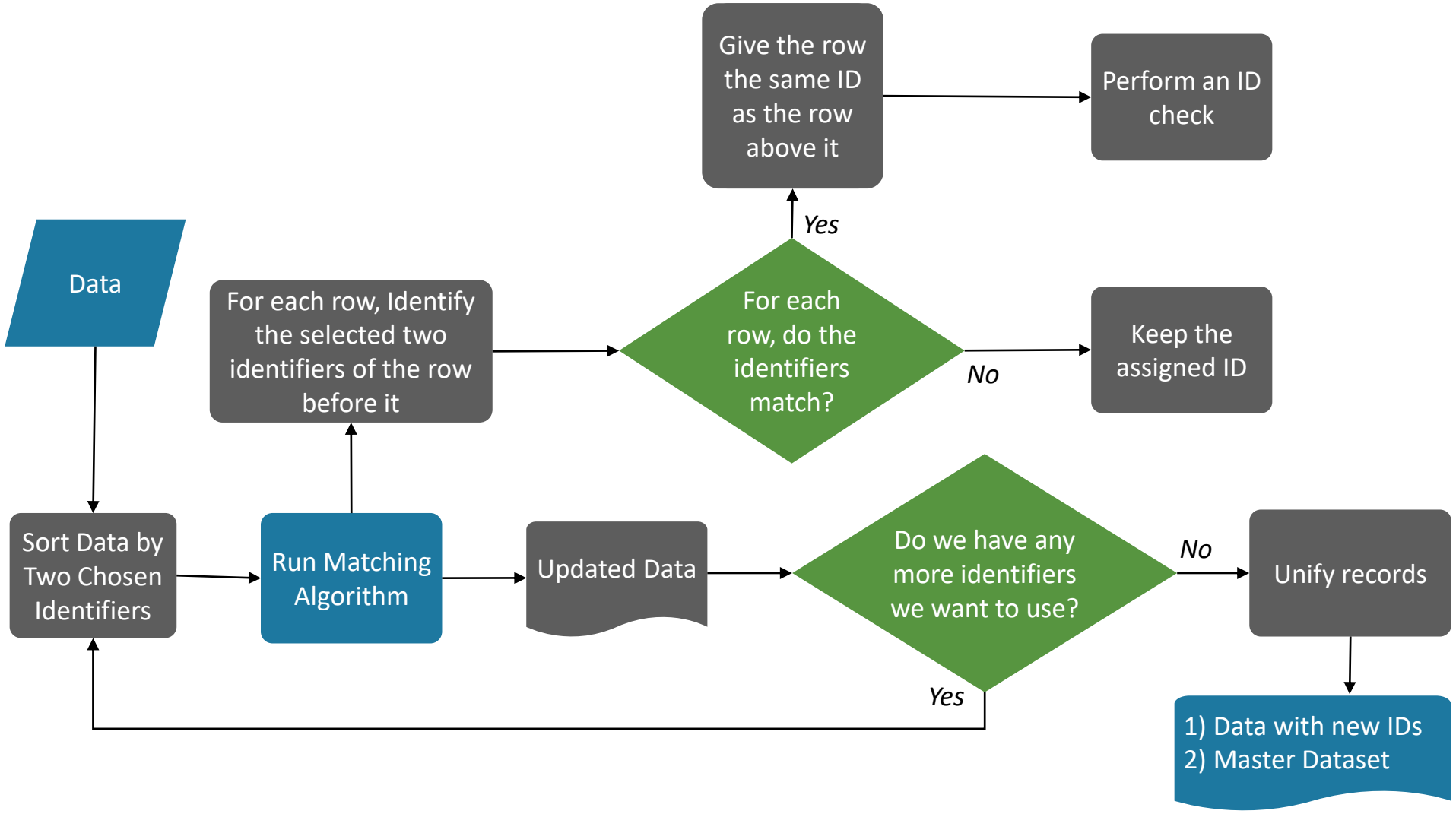
The team reviewed best practice matching methodologies to inform the proposed matching methodology for OIS

Sampling Approach

	Phase	Rationale
1	Sampled 12,000 record pairs	Large enough for all datasets and different combinations of identifiers
2	Sorted by four identifiers (anumber, Ident FIN, EID civ ID, name) and took 3,000 pairs from each	Not enough matching pairs if randomly sampled from 18 million records
3	Hand coded all 12,000 records as matches or non-matches	Easy for human to tell if pairs are a match or not
4	Entire group reviewed uncertain pairs and assigned together	Ensure consistency in uncertain matches
5	Dropped pairs where group could not determine match or non-match	All classifiers are binary, so no “unknown” value; only 12 cases affected
6	Divided into 70% training data and 30% testing data	Testing data to avoid overfitting


Introduction
Data
Approach
Business Rules
Probabilistic Models
Conclusion

Sorted Neighborhood Methodology



Fuzzy Name Matching

OIS Name Score

$$Name\ scoring\ function\ = \frac{\#\ of\ characters\ in\ common}{length\ of\ shortest\ name}$$

De La Cruz and Cruz: $\frac{4}{4} = 1.00$

Sanchez and Chavez: $\frac{5}{6} = 0.83$

OIS Business Rules

Names are a match if:

- First and last name scores are over the threshold
- Either first or last name score is over the threshold and the other is missing

Juan Diego Moralez and Juan Moralez: $\frac{4}{4}, \frac{7}{7} = 1$

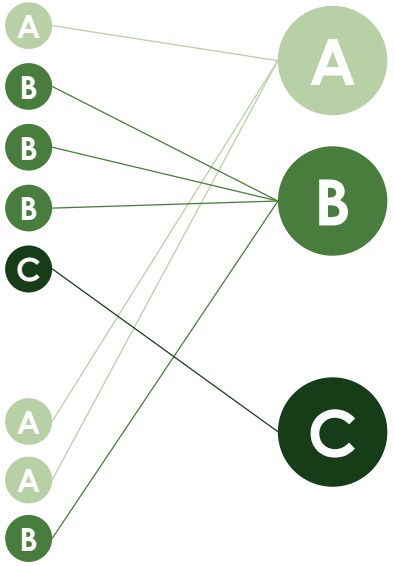
Rob Morales and Juan Moralez: $0, \frac{6}{7} = 0$

OIS developed business rules based on custom OIS scores of first and last name matches

Business Rules Match Criteria

Eight data sorts

1. Anumber and name (A)
2. Anumber and Ident FIN (B)
3. EID civ ID and anumber (B)
4. Ident FIN and anumber (B)
5. Name, country of citizenship, date of birth, and gender (C)
6. EID civ ID and name (A)
7. Ident FIN and name (A)
8. EID civ ID and Ident FIN (B)



Three ways to match

A — Numeric IDs
- At least two numeric identifiers match

B — Numeric ID and Name
- One numeric identifier matches
- The name-scoring function returns a 1 (threshold 50%) or the other numeric identifiers and name are missing

C — Biographic IDs
- The name-scoring macro returns a 1 (threshold 75%)
- Ident FIN, anumber, citizenship, gender, and birth date either match or are missing

Using business rules, data are sorted and matched eight times, seeking one of three ways to confirm a match

Business Rules Results

Results on Training and Testing Data

Accuracy: 0.8% Error Rate

Advantages

- Already implemented
- Code can be run on entire flow dataset in 1.5 hours using current computers

Disadvantages

- Annual dataset growth by 8-10 million records, increasing processing speed (1.5 hours for 38 million records)

Confusion Matrix of Results

green cells: correctly predicted
red cells: incorrectly predicted

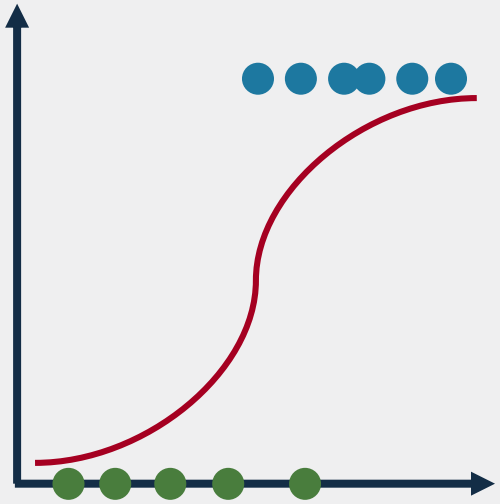
		Predicted Values: No Match	Predicted Values: Match	Error Rate
True Values	No Match	2,162	26	1.2% *(false positive)*
	Match	4	1,403	0.3% *(false negative)*

TOTAL ERROR RATE: 0.8%


Introduction
Data
Approach
Business Rules
Probabilistic Models
Conclusion

Logistic Regression Methodology and Results

Logistic Regression



A regression model with a categorical dependent variable

Results on Testing Data

Accuracy: 0.9% Error Rate

Advantages

- OIS can understand how the odds of a match increase for each piece of information
- Cutoff can be adjusted based on need
- Does not require first or last name
- Fast to apply: took 7 seconds on 5 million record test in R

Disadvantages

- Requires additional programming to implement

Confusion Matrix of Results

green cells: correctly predicted
red cells: incorrectly predicted

		Predicted Values: No Match	Predicted Values: Match	Error Rate
True Values	No Match	2,163	25	1.1% *(false positive)*
	Match	8	1,399	0.6% *(false negative)*

TOTAL ERROR RATE: 0.9%

Logistic Regression Cutoff Analysis

Match on Variables	Outcome	Probability
ANUM, BIRTH_DT, COUNTRY, GENDER	Match	99.9%
ANUM, BIRTH_DT, COUNTRY	Match	99.8%
ANUM, BIRTH_DT, GENDER	Match	98.4%
ANUM, COUNTRY, GENDER	Match	91.1%
ANUM, COUNTRY	Match	82.1%
BIRTH_DT, COUNTRY, GENDER	Match	76.4%
BIRTH_DT, COUNTRY	Match	59.3%
---------- 50 % Cutoff ----------		
ANUM,GENDER	No Match	37.6%
ANUM	No Match	21.3%
BIRTH_DT, GENDER	No Match	16.0%
COUNTRY, GENDER	No Match	3.0%
COUNTRY	No Match	1.4%
GENDER	No Match	0.2%
No Matches	No Match	0.1%

Naïve Bayes Methodology and Results

Naïve Bayes

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

A probabilistic classifier based on applying Bayes' theorem with (naïve) independence assumptions between the features

Results on Testing Data

Accuracy: 6.6% Error Rate

Advantages

- Able to easily extract conditional probabilities

Disadvantages

- Higher error rate than business rules and other models
- Very slow to apply: Over 20 minutes for 5 million record test

Confusion Matrix of Results

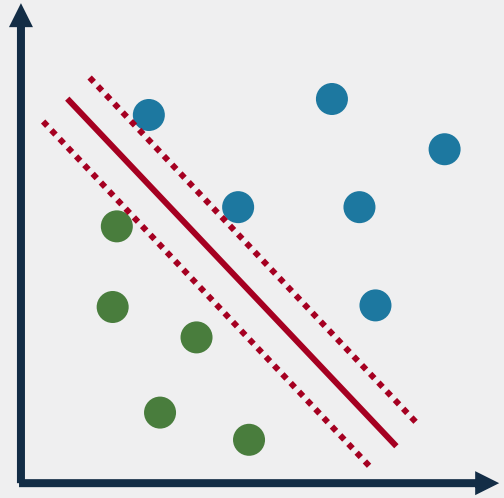
green cells: correctly predicted
red cells: incorrectly predicted

		Predicted Values: No Match	Predicted Values: Match	Error Rate
True Values	No Match	2,086	102	4.7% *(false positive)*
	Match	135	1,272	9.6% *(false negative)*

TOTAL ERROR RATE: 6.6%

SVM Methodology and Results

Support Vector Machine (SVM)



A model mapped so that the separate categories are divided by a clear gap that is as wide as possible

Results on Testing Data

Accuracy: 0.9% Error Rate

Advantages

- Low error rate

Disadvantages

- Slower than other methods: 118 seconds on 5 million record test
- Methodology is difficult to explain

Confusion Matrix of Results

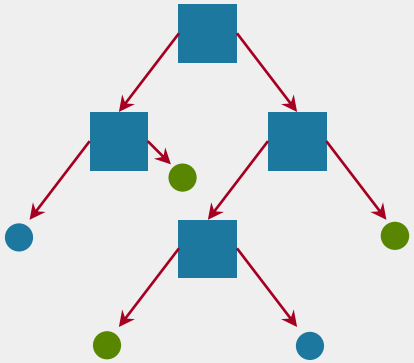
green cells: correctly predicted
red cells: incorrectly predicted

		Predicted Values: No Match	Predicted Values: Match	Error Rate
True Values	No Match	2,161	27	1.2% *(false positive)*
	Match	4	1,403	0.3% *(false negative)*

TOTAL ERROR RATE: 0.9%

Decision Tree Methodology and Results

Decision Tree



A decision support tool that uses a tree-like model of decisions and their possible consequences

Results on Testing Data

Accuracy: 0.9% Error Rate

Advantages

- Clear and transparent methodology
- Fast: 12 seconds on 5 million-record test
- Sorting only depends on two variables

Disadvantages

- 0.015% of records in flow dataset are missing both anumber and birth date
- 20% of records in flow dataset with anumbers do not have birth date

Confusion Matrix of Results

green cells: correctly predicted
red cells: incorrectly predicted

		Predicted Values: No Match	Predicted Values: Match	Error Rate
True Values	No Match	2,161	27	1.2% *(false positive)*
	Match	4	1,403	0.3% *(false negative)*

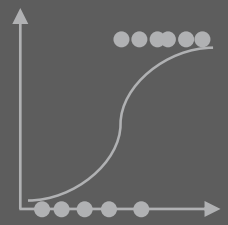
TOTAL ERROR RATE: 0.9%


Introduction
Data
Approach
Business Rules
Probabilistic Models
Conclusion

Conclusion

Five Tested Matching Techniques

Logistic Regression



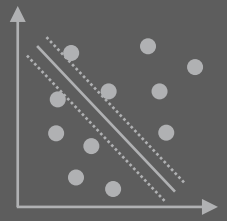
Business Rules

- A Numeric IDs
- B Numeric ID and Name
- C Biographic IDs

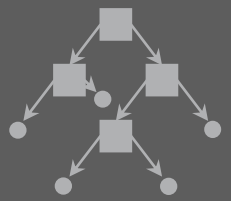
Naïve Bayes

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

SVM

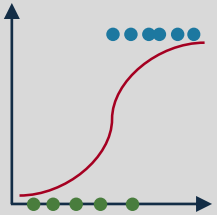


Decision Tree



Recommendation and Next Steps

Logistic Regression



- Low error rate
- Does not require first or last name
- Allows for a variable cutoff if desired

Next Steps

1. Implement logistic regression equation in SAS
2. Apply matching to OIS datasets for analysis

After testing five matching techniques, the team recommends using the **Logistic Regression** matching technique

References

Ai Pei, Shirley Ong. "A Comparative Study of Record Matching Algorithms." Master's thesis, University of Edinburg, Scotland, 2008. 2008. https://www.inf.ed.ac.uk/publications/thesis/online/IM080663.pdf.

Bilenko, M., R. Mooney, W. Cohen, P. Ravikumar, and S. Fienberg. "Adaptive name matching in information integration." IEEE Intelligent Systems 18, no. 5 (2003): 16-23. doi:10.1109/mis.2003.1234765.

Cochinwala, Munir, Verghese Kurien, Gail Lalk, and Dennis Shasha. "Efficient data reconciliation." Information Sciences 137, no. 1-4 (2001): 1-15. doi:10.1016/s0020-0255(00)00070-0.

Cohen, W., Ravikumar, P., & Fienberg, S. “A comparison of string metrics for matching names and records.” Kdd workshop on data cleaning and object consolidation, 2003.

Halwai, H., Mahajan, A., Pawar, N. “Rule Based Method or Entity Resolution”. International Journal for Scientific Research & Development, 2016.

Wilson, D. Randall. "Beyond probabilistic record linkage: Using neural networks and complex features to improve genealogical record linkage." The 2011 International Joint Conference on Neural Networks, 2011. doi:10.1109/ijcnn.2011.6033192.

Winkler, William E. “Methods for record linkage and Bayesian networks”. Statistical Research Division, US Census Bureau, Washington, DC, 2002.

Winkler, William E. "Methods for evaluating and creating data quality." Information Systems 29, no. 7 (2004): 531-50. doi:10.1016/j.is.2003.12.003.

Q&A

APPENDIX

Datasets - Detail

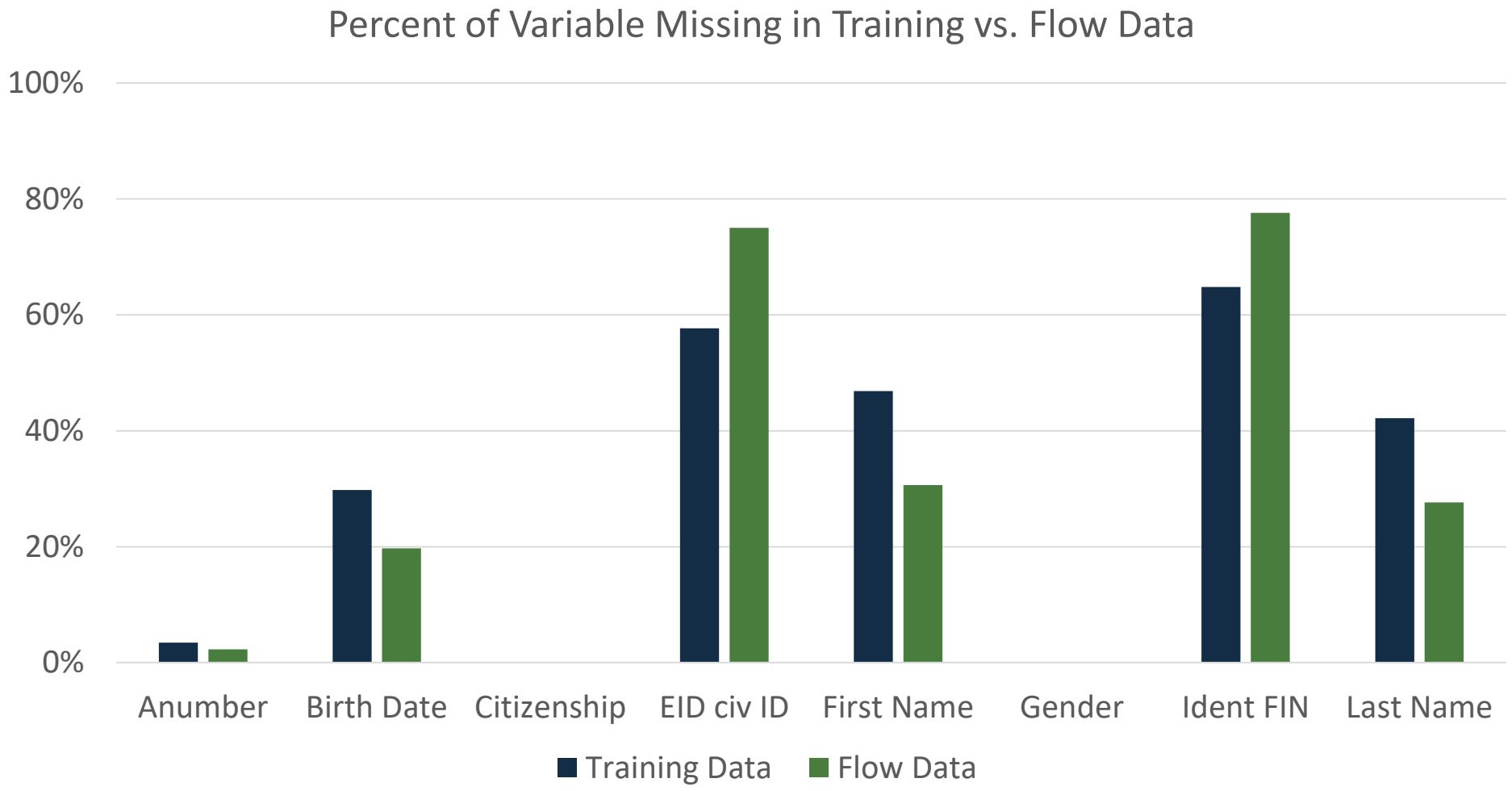
Dataset Name	Database	Database Name	Description of Data
Asylum data	RAPS	Refugees, Asylum, and Parole System	USCIS records of individuals who applied for asylum
Charging Documents Issued/ Notice to Appear (NTAs)	IIDS	ICE Integrated Decision Support	ICE records of individuals who have been issued a NTA
DACA	C3, ELIS	Electronic Immigration System	USCIS records of individuals who have been granted DACA
Defensive Asylum	CASE	Case Access System for EOIR	EOIR records of individuals who applied for defensive asylum
Detention (Bookins and Releases)	IIDS	ICE Integrated Decision Support	ICE records of individuals who have been booked in and out of detention
EAD	C3, ELIS	Electronic Immigration System	USCIS records of individuals who have been granted an Employment Authorization Document
EOIR	CASE	Case Access System for EOIR	EOIR records of proceedings and case
ERO Administrative Arrests	IIDS	ICE Integrated Decision Support	ICE records of individuals who have been arrested
HSI Administrative Arrests	ICM	ICE Investigative Case Management	ICE records of individuals who have been arrested
Legal Permanent Resident	C3, ELIS	Electronic Immigration System	USCIS records of individuals who have been granted Legal Permanent Resident Status
OFO Inadmissibles	Borderstat	Borderstat	CBP records of individuals who have been determined to be inadmissible
Removals and Returns	IIDS	ICE Integrated Decision Support	DHS records of individuals who have been removed or returned from the U.S.
USBP Apprehensions	Borderstat	Borderstat	CBP records of individuals who have been apprehended at the border

Variable Completeness in Flow Dataset

Percent missing by source component and variable

Component	First Name	Last Name	Anumber	Birth Date	EID civ ID	Ident FIN
U.S. Customs and Border Protection (CBP)	0.0%	0.0%	20.0%	0.0%	0.0%	17.5%
Executive Office for Immigration Review (EOIR)	100.0%	100.0%	0.0%	52.1%	100.0%	100.0%
Enforcement and Removal Operations (ERO)	20.5%	0.0%	0.6%	0.1%	0.0%	5.2%
Homeland Security Investigations (HSI)	12.3%	0.0%	26.7%	0.1%	100.0%	100.0%
United States Citizenship and Immigration Services (USCIS)	18.9%	18.9%	0.2%	18.9%	100.0%	100.0%

Training Dataset vs. Flow Dataset Comparison



What happens if there is a match in the middle of an ID group, Part 1?

Because OIS sorts our data multiple times by multiple identifiers, rows can be shuffled around and matches may not occur for every ID

In this example,

- Rows 1-6 have been previously matched to a person (ID 5), and row 7 is another person (ID 6)
- Rows 4-7 will match on anumber and eid_civ_id but rows 1-7 need to all point to person 6

Serial1	ANUM	EID_CIV_ID
5	123456	
5	123456	
5	123456	
5	123456	987654
5	123456	987654
5	123456	987654
6	123456	987654

Matching Algorithm Ran →

Serial2 (Post Match ID)	Serial1 (Pre Match ID)
5	5
5	5
5	5
6	5
6	5
6	5
6	6

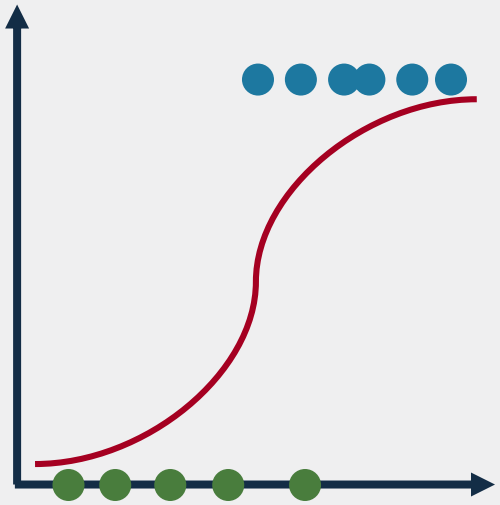
ID Check Ran →

Serial2 (Post Match ID)	Serial1 (Pre Match ID)
6	5
6	5
6	5
6	5
6	5
6	5
6	6

What happens if there is a match in the middle of an ID group, Part 2?

- The previous matching macro did not address the consistent handling of ID swaps. Because the team sorts the data multiple times, the IDs may not be properly updated each iteration
- The team implemented a hash solution to properly handle these situations
  - The team created a hash object based on a old_id -> new_id crosswalk
    - To save SAS computation time, the team ignored cases where old_id = new_id in the newly matched data set
  - For each record with a new_id, the team checked the hash table to see whether it was ever in the "old_id" position
  - If the team finds a row, then the team outputs the new_id
  - If the ID was ever changed, the team now moves that old_id position and checks again with the hash table
  - This continues a loop until the swap does not occur (or if the team searches over 1,000 times)
  - When the team stops matching, the team knows that the value currently stored as the old_id does not occur in the crosswalk table with any new_id assigned to it
- A hash solution is much faster than a do loop that iterates through the entire SAS data set
- 0.01% of observations looped over 1,000 times

Logistic Regression Detail



A regression model with a categorical dependent variable

Coefficients	Estimate	Std. Error	z value	P(>|z|)
Intercept	-7.0988	0.4331	-16.39	< 2e-16
anum_match1	5.7929	0.2244	25.819	< 2e-16
birth_dt_match1	4.643	0.2854	16.266	< 2e-16
country_match1	2.8311	0.4353	6.503	7.87E-11
gender_match1	0.7978	0.2303	3.464	0.000531

Naïve Bayes Detail

Naïve Bayes

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

A probabilistic classifier based on applying Bayes' theorem with strong (naïve) independence assumptions between the features

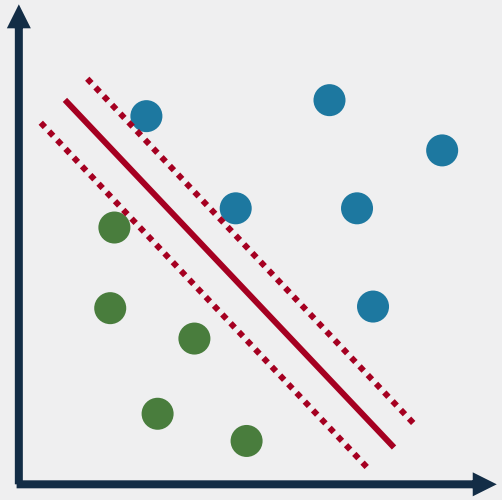
Conditional Probabilities from Naïve Bayes Model

Note: All probabilities are independent

Variable	Status	P(Variable Match) or Score Mean | True Match
Anumber	Match	97%
	Missing	3%
Birth Date	Match	87%
	Missing	10%
Citizenship	Match	99%
	Missing	52%
EID Civ ID	Match	52%
	Missing	34%
First Name	Score Mean	67%
	Missing	31%
Gender	Match	94%
Ident FIN	Match	55%
	Missing	44%
Last Name	Score Mean	75%
	Missing	20%

SVM Detail

Support Vector Machine (SVM)



A model mapped so that the separate categories are divided by a clear gap that is as wide as possible

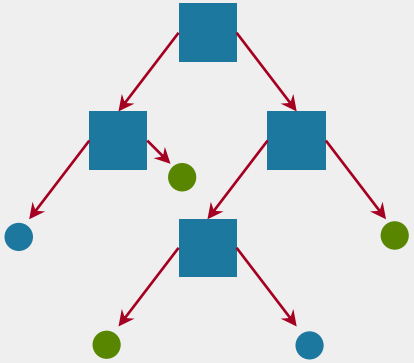
Output: 264 support vectors, using a radial svm-kernel (exp(-gamma*|u-v|^2),) gamma=0.059

Sample SV	Anum_match0	Anum_match1	Anum_miss1	Birth_dt_match 1	Birth_dt_miss1	First_name_com pged	Last_name_com pged	Country_match1
1073	1	0	1	1	0	0.53764370	-0.57105405	1
7147	0	1	0	0	1	1.43282114	3.33431737	1

Sample SV	Eid_civ_id_ma tch1	Eid_civ_id_mi ss1	First_name_m iss1	First_name_sc ore	Gender_match 1	Ident_fins_ma tch1	Ident_fins_mis s1	Last_name_mi ss1	Last_name_sc ore
1073	1	0	1	-0.7528986	1	1	0	0	1.23124546
7147	0	1	1	-0.7528986	0	0	1	1	-0.89523801

Decision Tree Detail

Decision Tree



A decision support tool that uses a tree-like model of decisions and their possible consequences

OIS Decision Tree