This document addresses some of the frequently asked questions regarding SPSS Modeler Premium Entity Analytics (EA) ranging from what is EA through technical details on how to use the product.

View more content in this series

Introduction

Purpose of document

This document addresses some of the frequently asked questions regarding IBM® SPSS® Modeler Premium Entity Analytics (EA) ranging from “What is EA?” through technical details on how to use the product.

Applicability

This document is applicable to the IBM SPSS Modeler Premium Entity Analytics product as a whole and is intended to be version indifferent.

Exclusions and exceptions

The information contained in this document may evolve over time as features change.

Assumptions

Entity Analytics is a premium component of IBM SPSS Modeler. Users of this document should have a base understanding of SPSS Modeler, Entity Analytics, and the data they will be accessing.
**EA frequently asked questions**

**What is IBM SPSS Modeler Premium Entity Analytics (EA)?**

Entity analytics allows an organization to combine diverse data sources to determine which entities are the same and can be resolved, even when the entities do not share any key values such as the same customer or account number. This capability creates more accurate models which in turn generate better business outcomes. Entity analytics shall help organizations more accurately score risk (e.g., reducing both false positives and false negatives in anti-money laundering programs) and improve score opportunity (e.g., de-duplicating customer files thus able to assess whole value).

For additional information on increasing accuracy of models using EA, see the Resources section at the end of this article for a link to the IBM Redbook titled **Using Entity Analytics to Greatly Increase the Accuracy of Your Models Quickly and Easily**.

**What matching technologies does EA use?**

Entity Analytics uses a set of **Principle Based Matching Rules** based on **Feature Behavior**. So whether the **entities** you are trying to match are persons, organizations, vehicles, or objects, the rules need not change.

Nearly every kind of entity has associated identifiers such as known locations, physical attributes, and names. These are called features within EA which comes pre-configured with the key features needed for matching persons, organizations, vehicles, and vessels so there is not much overhead to start matching these kinds of entities right out of the box. See Table 1 for a list of some examples of features that are supplied with EA.

**Table 1: Pre-configured features settings**

<table>
<thead>
<tr>
<th>FEATURE</th>
<th>FREQUENCY</th>
<th>EXCLUSIVITY</th>
<th>STABILITY</th>
</tr>
</thead>
<tbody>
<tr>
<td>NAME</td>
<td>NAME</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>DOB</td>
<td>FM</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>DOD</td>
<td>FM</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>GENDER</td>
<td>FM</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>ADDRESS</td>
<td>FF</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>PHONE</td>
<td>FF</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>SSN</td>
<td>F1</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>SSN_LAST4</td>
<td>FM</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>DRLIC</td>
<td>F1</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>PASSPORT</td>
<td>F1</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>TAX_ID</td>
<td>F1</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>GOVT_ID</td>
<td>F1</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>LOGIN_ID</td>
<td>F1</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>EMAIL_ADDR</td>
<td>FF</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>
Many of these features are locked and cannot be changed in Modeler as some of their settings, such as how they are standardized and compared, are not exposed in the user interface. Use the existing features as-is whenever possible. They can also be used as a guide when setting up new types of entities.

**How are feature behaviours set?**

If new entities or new features need to be added, three questions are required to be answered to define their behaviour for the matching rules. As shown in Table 1, the three questions relate to **Frequency**, **Exclusivity**, and **Stability**. When answering these questions keep in mind that there are always exceptions. For instance, only one person should be associated with a Social Security Number and it usually does not change throughout their lifetime. In rare situations however, such as in identity theft, new numbers can be issued. When answering the below questions on the creation of a new feature, answer for how a feature normally works.

**Frequency** - How many entities should have this feature value?

- **One (F1)** - Only one entity should have this value. These are usually ID numbers and account numbers.
- **Few (FF)** - Only a few entities should have this value. These are usually methods of contacting the entity, such as their address, phone, or email.
- **Many (FM)** - Many entities can have this value, such as dates of birth and physical attributes.
- **None** - This is essentially for internal use features that should not affect resolution in any way.
- **Name** - This is reserved for person or organization names only. Never assign this to any feature other than the existing NAME feature that ships with EA with. If some other kind of entity has a “name” then use the existing NAME feature.

**Exclusivity** - Should an entity only have one of these?

Exclusive features help break what are otherwise weak matches. For instance if only a name and address match, then a different Date of Birth (DOB), ID number, or physical attribute such as gender should break the match.

- **Yes** - For ID numbers and physical attributes such as DOB, gender, etc.
- **No** - For everything else.

**Stability** - Should it be the same throughout the entity's entire life?

Stability is like an exclamation point on Exclusivity and should not be Yes if Exclusivity is not also Yes. Stable features can break even strong matches. For instance if two entities share a close name and account number but have a different DOB, it may be a false positive.

- **Yes** - For very reliably entered features such as government issued IDs and DOBs.
- **No** - For everything else.
Be certain when setting stability to Yes. While place of birth should also not change throughout an entity’s life, it is not a reliably entered field and would break too many matches simply because it is rarely standardized and hard to match.

**How are entities resolved?**

When resolving entities, the following steps are taken:

1. The incoming entity’s features are compared against all others in the repository resulting in a short list of possible matching entities referred to as **candidates**.
2. All of the candidate record’s features are compared with the incoming features utilizing special scoring returns described below.
3. The scores are then summarized into behaviour categories to determine how many features in that category are considered to be the same or at least close.
4. The following set of resolution rules is applied to each candidate:
   a. If a **resolve rule** is hit, the incoming entity is considered to be the same as the candidate entity and is given the same unique ID. If no resolve rules are hit, a new unique ID is created for the incoming entity.
   b. Candidates that hit a **relationship rule** are related to the incoming entity’s assigned unique ID. This is known as a **derived relationship** and is only available in the unleashed version.
   c. Candidates that do not hit any resolve or relate rules are simply discarded.

**What are the actual resolution rules EA uses?**

The resolution rules and scoring are as follows. They are applied top down and only the first matching rule is applied. There is an overall rule setting to allow the rules to behave in a **Default** (best practice), **Aggressive** (more optimistic), or **Conservative** (more pessimistic) behaviour.

**Table 2: Resolution rules EA uses**

<table>
<thead>
<tr>
<th>#</th>
<th>What should match</th>
<th>Score</th>
<th>Aggressive</th>
<th>Default</th>
<th>Conservative</th>
</tr>
</thead>
</table>
| 101 | SF1_PNAME_CFF_CSTAB  
SAME F1 (an ID or account #)  
PARTIAL NAME (at least first name)  
CLOSE FF (an address, phone, or email)  
CLOSE STABLE (a DOB or other feature marked stable) | 10    | Resolved Same | Resolved Same | Resolved Same |
| 102 | SF1_PNAME_CFF_DSTAB  
SAME F1 (an ID or account #)  
PARTIAL NAME (at least first name)  
CLOSE FF (an address, phone, or email)  
DIFFERENT STABLE (a DOB or other feature marked stable is different) | 8     | Resolved Same | Resolved Same | Related / Possibly Same |
| 103 | SF1_PNAME_CFF  
SAME F1 (an ID or account #)  
PARTIAL NAME (at least first name)  
CLOSE FF (an address, phone, or email)  
NO STABLE (no DOB or other feature marked stable was found) | 9     | Resolved Same | Resolved Same | Resolved Same |
| 104 | SF1_PNAME_CSTAB  
SAME F1 (an ID or account #)  
PARTIAL NAME (at least first name)  
CLOSE STABLE (a DOB or other feature marked stable) | 10    | Resolved Same | Resolved Same | Resolved Same |
<table>
<thead>
<tr>
<th>ID</th>
<th>Rule Description</th>
<th>Count</th>
<th>Resolution 1</th>
<th>Resolution 2</th>
<th>Resolution 3</th>
</tr>
</thead>
</table>
| 105 | SF1_PNAME_DSTAB
SAME F1 (an ID or account #)
PARTIAL NAME (at least first name)
DIFFERENT STABLE (a DOB or other feature marked stable is different) | 8     | Resolved     | Resolved     | Related / Possibly Same |
| 106 | SF1_PNAME
SAME F1 (an ID or account #)
PARTIAL NAME (at least first name) | 9     | Resolved     | Resolved     | Resolved               |
| 107 | PNAME_MSTAB_CFF
PARTIAL NAME (at least first name)
MULTIPLE STABLES (multiple stable features match, like DOB and Last4SSN)
CLOSE FF (an address, phone, or email) | 10    | Resolved     | Resolved     | Resolved               |
| 108 | PNAME_MSTAB
PARTIAL NAME (at least first name)
MULTIPLE STABLES (multiple stable features match, like DOB and Last4SSN) | 9     | Resolved     | Resolved     | Resolved               |
| 109 | CNAME_CFF_CSTAB
CLOSE NAME (both first and last name are at least close)
CLOSE FF (an address, phone, or email)
CLOSE STABLE (a DOB or other feature marked stable) | 8     | Resolved     | Resolved     | Resolved               |
| 110 | CNAME_CFF_DEXCL
CLOSE NAME (both first and last name are at least close)
CLOSE FF (an address, phone, or email)
DIFFERENT EXCLUSIVE (a DOB, gender or other feature marked exclusive is different) | 6     | Related / Possibly Same or Know each other | Related / Possibly Same or Know each other | Related / Possibly Same or Know each other |
| 111 | CNAME_CFF
CLOSE NAME (both first and last name are at least close)
CLOSE FF (an address, phone, or email) | 7     | Resolved     | Related / Possibly Same or Know each other | Related / Possibly Same or Know each other |
| 112 | SF1_CFF_CSTAB
SAME F1 (an ID or account #)
CLOSE FF (an address, phone, or email)
CLOSE STABLE (a DOB or other feature marked stable) | 8     | Resolved     | Resolved     | Resolved               |
| 113 | SF1_NCONF
SAME F1 (an ID or account #)
NO_CONFLICTS (no conflicting name or exclusive feature) | 7     | Resolved     | Resolved     | Resolved               |
| 114 | MSTAB_NCONF
MULTIPLE STABLES (an ID or account #)
NO CONFLICTS (no conflicting name or exclusive feature) | 7     | Resolved     | Resolved     | Resolved               |
| 115 | SF1
SAME F1 (an ID or account #) | 5     | Related / Possibly Same or Know each other | Related / Possibly Same or Know each other | Related / Possibly Same or Know each other |
| 116 | SNAME_SSTAB
SAME NAME (both first and last name are at the same)
SAME STABLE (a DOB or other feature marked stable) | 6     | Related / Possibly Same | Related / Possibly Same | Related / Possibly    |
| 117 | SURNAME_CFF
SUR NAME (surname or last name only)
CLOSE FF (an address, phone, or email) | 5     | Related / Possibly Know each other | Related / Possibly Know each other | Related / Possibly Know each other |
| 118 | CFF
CLOSE FF (an address, phone, or email) | 4     | Search only  | Search only  | Search only            |
| 119 | CNAME | 2     | Search only  | Search only  | Search only            |
When reviewing the rules in Table 2 above, some rules can be grouped into similar concepts. The following will describe some of these groupings.

101-103: The first rule checks for a confirming DOB, the second rule checks for a conflicting DOB, and the third rule is for when there is no DOB.

104-106: If you look carefully, these three rules match the first three rules, except that there is no address, phone or email that also matches. The score is then a little bit lower and the Default and Conservative rule sets will trigger a possible match relationship rather than a resolution if the DOB or other stable element is missing (see rule 105).

107-108: These two rules introduce the concept that if there are no identifiers (F1s), then multiple stable elements are a strong enough indication of a match. For instance, the combination of DOB and the last four digits of a Social Security Number (SSN) are shared by very few people. Not quite an F1, but still a very good indication of a match.

109-111: Here are the trio of rules again except name and an FF is the primary key because we have neither a matching F1 nor multiple stable features. Now we are not so quick to resolve and the middle rule looks for any conflicting EXCLUSIVE feature to break the match.

112: This rule is for when the name does NOT match but there is overwhelming other evidence that it is the same entity.

113-114: This set of rules exist for when there is no matching name:

- Some types of entities such as vehicles, vessels, and devices simply don’t have a name to match with.
- Even such as transactions as persons and organizations recordsets only carry account number (an F1) so that is all that can be used to match.

115: This rule indicates that you have two different entities sharing an identifier. We know there are other conflicts between the two records otherwise they would have hit rule 113 and been resolved. This can indicate identity theft or spouses using the same account numbers.

116: While name and DOB is really not strong enough to resolve two entities as the same, a possible match relationship is created for it.

117: A matching home address alone is very often a false positive if it is not known that the entities lived at the same address at the same time. However, if two entities share the same last name and address, that can indicate they are possibly from the same family, so a relationship is created.

118-119: These two rules neither resolve nor relate as the match just isn’t strong enough to consider that they are the same entity or at least know each other. But they are only used by the streaming EA node which is most often used to find possibly matching records at a given moment in time. For instance, rule 119 would fire if you searched for someone by name alone.
Which resolution rule setting should I pick?

Resolving entities can be looked at a couple of different ways.

1. You are simply trying to determine if two records match.
2. You are trying to determine if the two entities really are the same.

To be really certain that two entities are the same, it generally requires ID numbers (Frequency One) and DOBs (Exclusive/Stable) features to be really certain that two entities are the same - we call these confirming features. When you have sparse data sets that have only name and address (Frequency Few) features, you may want to consider two records match even without the confirmation of an ID number or DOB. Such is often the case when trying to match company data.

Here is a guideline on when to use which setting,

- **Default** is a pretty good balance when you have:
  - Identifiers or DOB (exclusives/stables)
  - Confirms usually override conflicts
- **Conservative** requires:
  - Identifiers or DOB (exclusives/stables)
  - Conflict stable breaks match
- Only use **Aggressive** with sparse data:
  - Will match on name and address alone
  - Unless a conflicting exclusive/stable

The Default and Conservative rule settings are very close in the results they produce. The Aggressive rule set can return many more matches.

How are features scored?

- Names are matched using IBM Global Name Recognition which uses a linguistic based approach founded on more than 20 years of research and use.
- Dates such as DOB are matched using a proprietary algorithm that looks for such things as month/day swaps, 1 year off or 1 day off.
- ID numbers and other strings are matched with a Levenshtein edit distance algorithm to determine closeness.
- Addresses are matched with a proprietary algorithm that looks for specific parts of the address to match, such as street number, street name, unit number, city, and postal code.

What data cleansing should be performed before sending data to EA?

While EA attempts to match through dirty data, some data cleansing and quality checking should be performed prior to sending data through EA. Here are some best practices for getting data ready for EA.

- **Names** - Organization names should be mapped the ORG_NAME feature element while individual names should be mapped to the SUR_NAME, GIVEN_NAME and MIDDLE_NAME
feature elements. There is also a FULL_NAME feature element that can be used if you just can't tell the difference.

Sometimes the source record can have both an organization name and an individual name on it or even a primary name and an aka or nick name. It is important to distinguish between the names by specifying a usage type as well. For instance, you might specify the ORG_NAME as primary and the SUR_NAME/GIVEN_NAME combination as secondary.

• **Addresses** - It is best to perform address standardization prior to EA especially for international addresses. In fact, you can map directly to the address components that are used for matching - street number, street name, unit number, city, and postal code. If you cannot pre-parse the address field (ADDR1 for example), EA will attempt to parse the field into the specific feature elements (street number, street name, unit number) automatically. Ideally you should map the following address components,
  • ADDR1 or (STR_NUM, STR_NAME, UNIT_NUM)
  • CITY
  • POSTAL_CODE
• **Generic values** - While EA detects when values are overused, it is best to prevent these from entering the system to begin with. Here are some things to look for and clean before sending to EA,
  • Account or ID Numbers with the value **000000000** or even **n/a**. Often you can get a list of generic values by putting the source field through an aggregate node performing a count, looking for values used more than 10 times. **An ID number is supposed to have the frequency of one.**
  • DOB’s that are incomplete (not valid dates) or are overused. All too often **1/1/1900** gets to be the default DOB either due to data entry practices (filling a required field with any value) or the data extract you received was flawed in some way.
Resources

- IBM Redbook - Using Entity Analytics to Greatly Increase the Accuracy of Your Models Quickly and Easily
About the author

Steve Schormann

Steve Schormann has been with IBM in various positions for 32 years. Most recently in the Identity Insight development group as senior performance engineer. Prior to this Steve was in the the DB2 development group in Toronto, much of that time working with the performance group in IBM DB2 to achieve some world record performance benchmarks. Steve then was involved in database compatibility efforts within DB2, developing tools and assisting ISVs and customers in their migrations to DB2.

© Copyright IBM Corporation 2014
Trademarks
(www.ibm.com/developerworks/ibm/trademarks/)