Record Linkage with College Football

Haven King-Nobles, Jeremy Abramson
Information Sciences Institute
hkingnob@oberlin.edu, abramson@isi.edu

Motivation

- Many entities *do not* have unique identifiers that are shared across different data sets.
  - This complicates the process of aggregating data without creating duplicates.
- College football has several different databases going back more than 10 years.
  - Research question: How can a program link X in one data set to Y in another?
- Our contribution:
  - Record linkage tool to aggregate college football team and player data from different data sets.
  - Database of linked player data.
  - Some data analytics – Determining which counties produce the best recruits.

Challenges and Results

Challenges: Not all teams or players are referred to identically (e.g., abbreviations, misspellings). Sometimes different players share identical names. Identifiers (e.g., school) change as players grow older.

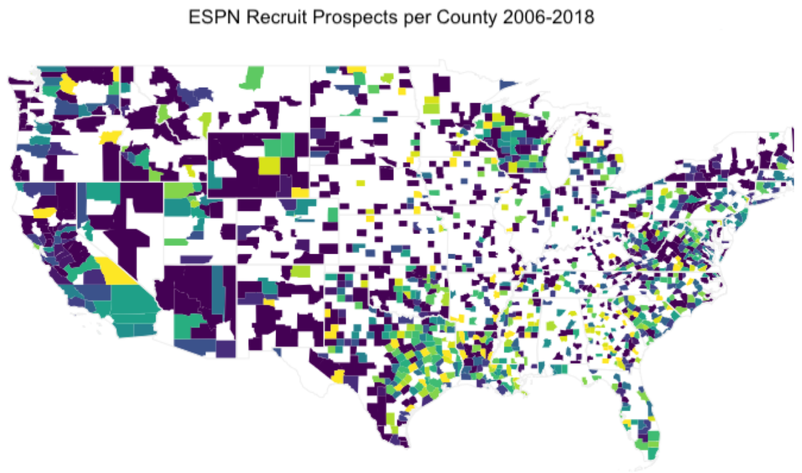
Initial Results (with team record linkage):

Canonical Team Name	NCAA	ESPN	247Sports
Eastern Michigan	Eastern Mich.	E Michigan	Eastern Michigan
Florida Atlantic	Fla. Atlantic	FAU	Florida Atlantic
Florida	Florida	Florida	Florida
Florida International	<no confirmed match>	FIU	Florida International
Florida State	Florida St.	Florida State	Florida State

Process

1. **Gather Data**
   - Scrape databases of sites like ESPN, NCAA, and 247Sports.
   - Use of Beautiful Soup library, JSON data, and headless browsing.
2. **Record Linkage**
   - Determine which records refer to the same entities, and which do not.
   - Use of distance metrics (e.g., Levenshtein and Jaro-Winkler Similarity).
3. **Data Analysis**
   - Determine accuracy of record linkage and use it to draw conclusions of the aggregate data of multiple data sets.

Which counties produce the best football recruits?



Top 3 Counties for Highest Ranked Recruits (with >5 recruits):

1. Long (GA) – 6 recruits
2. Gage (NE) – 6 recruits
3. Manatee (FL) – 313 recruits

The choropleth above displays every county that has produced NCAA football recruits over the last 12 years. Colder colors (yellow and green) represent counties with the greatest proportion of highly-ranked (3-, 4-, and 5-star) recruits, while warmer colors (dark purple and blue) represent counties with the lowest proportion of highly-ranked recruits. Note that many of the lower-population counties that scored high did so for having produced a very small number of highly-ranked recruits.

Record Linkage Methods

Approach: Concatenate different string identifiers together and compare using string distance metric (i.e., Jaro-Winkler similarity – see below).

Example: For team record linkage, we compared on the team name concatenated with the team mascot, which was repeated twice (because mascots tend to be abbreviated less than team names). For example, BYU, which is represented in one source as *BYU CougarsCougars*, matches most closely to *Brigham Young CougarsCougars* in another source.

Jaro-Winkler similarity: minimum number of single-character transpositions necessary to transform one string to another, with the prefix (in our case the first letter) given greater weight.

Evaluation Metrics – Accuracy and Usability

Accuracy and Usability

- High confirmation rate for team record linkage.
- All matches below a given Jaro-Winkler threshold score outputted to be manually reviewed.
- Databases of hundreds of teams and hundreds of thousands of players.
- Results outputted to CSV and Sqlite3 tables.

Percent Confirmed Matches and False Matches per Source

	NCAA	ESPN	247 Sports	Rivals*
Confirmed Matches	95%	94%	95%	75%
False Matches	2%	2%	2%	20%

*Linkage involving the Rivals data set performed significantly poorer due to that set's lack of mascot information and slightly different set of teams.

Future Work

- Player data record linkage.
- Analysis of linked data.
- E.g.: Comparing recruiting data and college data, are there any correlations between percentage change of certain performance metrics (e.g., 40-yard dash time) and college team?

If interested contact Haven King-Nobles (hkingnob@oberlin.edu)
Work performed under REU Site program
supported by NSF grant #1659886

USC Viterbi School of Engineering
Information Sciences Institute