Record linkage using supervised machine learning

Springboard Data Science Career Track: Capstone 1

Maya Bhat-Gregerson

March 9, 2020

Overview

- Linkage of administrative data yields great value to public health agencies and health researchers.
- The capstone is a pilot project to see if linkage of birth and death records either manually or through rough SQL queries can be replaced by supervised machine learning methods.
- In comparison with manual linkage of these records, the logistic regression and random forest classifiers both yielded improvement in terms of reducing time and the need for manual review in linking infant birth and death records.
- The algorithms will need to be optimized further, but show promise for broader application.

Background

- Washington State Department of Health, Center for Health Statistics produces linked birth-death record data sets for use by public health researchers.
- Limited to linkage for decedents under the age of 1 year. This data set is also known as the infant birth-death linked file.
- Comprehensive linked administrative sets connecting birth, death, hospital discharge records, prescription monitoring program data, and other data sets do not currently exist.
- A linkage process bridging these data sets would provide a rich source of information for epidemiological research that could translate into targeted public health policy.

Overview of workflow

This is a high level view of the steps taken from data acquisition to evaluating the supervised learning classifiers

Acquisition	• Retrieve birth, death data using SQL script
Cleaning	• Standardize values, keep only linking variables
Prep/EDA	• Select variables for linking from birth data and death data based on EDA
Train classifiers	• Logistic Regression • Random Forest
Evaluate and compare	• Based on accuracy scores, ROC curves

Data acquisition

- Birth and death records (2016-17) retrieved from SQL database
  - Connection to SQL server and SQL query written within Python script
  - Restricted to variables likely to contribute to training classifiers including first and last names of infant and infant's parents, date of birth, sex, residence county.
- Labelled data: manually linked infant birth-death file for 2016-17

Data cleaning

- TEXT FIELDS:
  - Strip punctuation, white space
  - Standardize to lower case
- DATE FIELD:
  - Split into month, day, year
  - Format month, day, year fields as integers
- APPLY RESTRICTION CRITERIA:
  - Included only Washington State residents who died in Washington and whose mother's were Washington State residents. This assures that all records are accessible.
  - Less than 1 year old at time of death
  - Births, deaths occurred in 2016-17

Prepare and explore data (1)

- Creating a feature set to train the classifiers consists of two steps:

1) CREATE CANDIDATE PAIRS:

- Each infant death record (n=631) was paired up with each record in a random sample of death records (n=650).
  - There are approximately 58,000 deaths per year. Creating candidate pairs using all death records would create a huge computational load.

2) COMPUTE SIMILARITY SCORES

- Apply a variety of functions to compute similarity scores for corresponding fields in birth records and death records e.g. Jaro-Winkler string similarity score based on comparing first name of infant in death record with first name of infant in birth record etc.
- Resulting data frame consists of a series of similarity scores (one row of scores for each pair of records compared)

Prepare and explore data (2)

- EXPLORATORY DATA ANALYSIS
  - Iterative process throughout model training
  - Explored similarity score for each pair of variables by match status (matching records vs. non-matching records)
  - Variables that didn't show separation of similarity scores by match status were considered for elimination in subsequent rounds of training classifiers
  - Looked for correlation between variables to remove highly correlated variables

Train classifiers

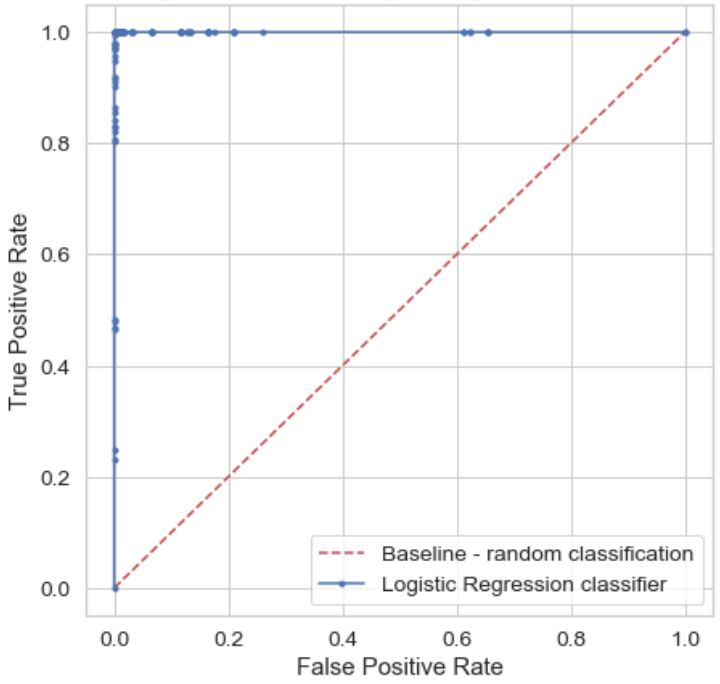
- Logistic Regression
  - using similarity scores for all variables in data set (names, sex, date of birth, county of residence, phonetic encoding of names)
- Random Forest
  - 100 trees
  - Used bootstrapping
  - No maximum set for number of features to consider for splitting node
  - No maximum for number of levels
  - This classifier took longer to finish predictions compared with logistic regression.

Evaluation and comparison of models

Logistic Regression

	No Match	Match
Precision	0.99	0.98
Recall	0.99	0.95
F1 score	0.99	0.96

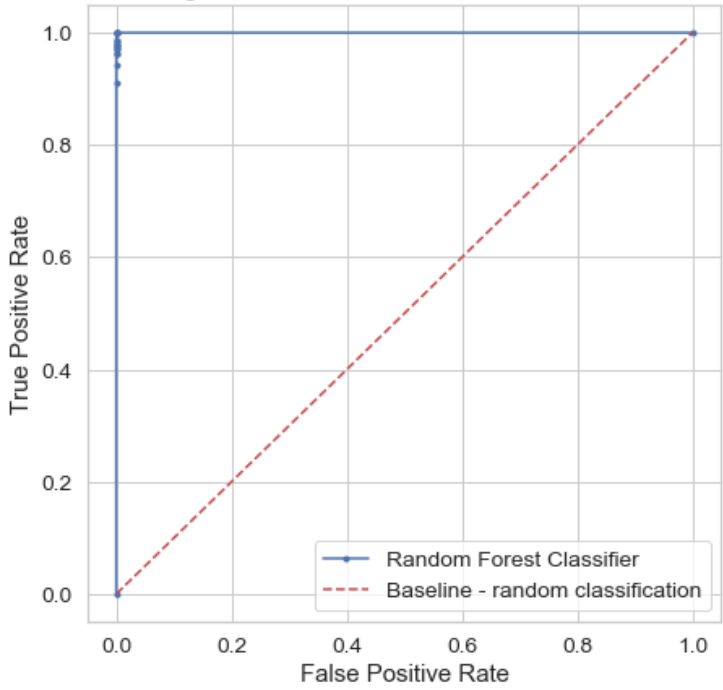
Fig 2. ROC curve for logistic regression classifier



Random Forest

	No Match	Match
Precision	0.99	0.98
Recall	0.99	0.97
F1 score	0.99	0.98

Fig 4. ROC curve for random forest classifier



Next steps

- Optimization
  - Hyper-parameter tuning with grid search and cross validation
- Validate models
  - Apply models to 2018 and 2019 infant birth and death data to see if similar accuracy can be achieved.
  - If not – explore and address overfitting.
- Apply final models to other data
  - Other death – birth linkage (all ages)
  - Link birth or death to other data sets e.g. hospital discharge where there may not be a one-to-one match.