Predictive Analysis on Recidivism

For safer and more productive communities.

Team members : James Avery, Robert Hanlon, Cindy Reznikov

Recidivism - Latin recidiv(us) - relapsing

- ❖ Overview
- ❖ Data Gathering and Cleaning
- ❖ Machine Learning Models
  - ➢ Logistic Regression
  - ➢ SVM
  - ➢ Naive Bayes
- ❖ Tableau Visualizations

Overview

To increase public safety and reduce costs, many US states are evaluating their criminal justice policies and procedures. They are enacting and evaluating ongoing performance measures. In order to reduce recidivism, it is important to:

1. Use multiple measures of success (i.e. track for at least 3 yrs after release, desistance (never return to prison), time to failure, crime severity (initial/subsequent), etc.)
2. Consistent, accurate and timely collection of data
3. Analyze the underlying composition of the population
4. Broad dissemination of the analysis to all 3 branches of state government

Data gathering and cleaning:

- 3-year recidivism for offenders released from prison in Iowa (2013-2018) https://dev.socrata.com/foundry/data.iowa.gov/mw8r-vqy4 (26,020 records)

  Iowa Department of Corrections
- Split out race and ethnicity
- Dropped or filled in nulls in certain columns
- Removed leading and trailing spaces in column names

Machine Learning: Logistic Regression

- StandardScaler & MinMaxScaler from Scikit-Learn
- RFE - to improve feature selection
- GridSearchCV - hyperparameter tuning
- Ran multiple tests adjusting features based on RFE and GridSearchCV parameters
- Final result: 67%

Machine Learning: SVM

- Used Pandas to get dummies, SVM only interacts with numerical data.
- StandardScaler from Scikit-Learn
- SVM from Scikit-Learn

Initial results: 66%

- SVM is great for categorizing numerical data but doesn’t seem to be utilized as much with categorical data.
- Best used with numerical data such as the iris data and measurements datasets and does not work well if any null values are present.

Final results: 66%

Machine Learning: Naive Bayes

- MinMaxScaler from Scikit-Learn

Initial results: 36% accuracy -- not good

- Possible implication is that features are not orthogonally independent as would be ideal for use of the Naive Bayes model.
- Pro: faster than the other two models to train.
- "[...] no matter how strong the dependences among attributes are, Naive Bayes can still be optimal if the dependences distribute evenly in classes, or if the dependences cancel each other out." - Dr. Harry Zhang, "The Optimality of Naive Bayes"
    - It is likely that our dependencies do not distribute evenly, or do not cancel each other out; hence, a suboptimal result for our data and feature selection.

Thank you!