Machine Learning Student Projects
Public Safety Lab
New York University

Gregory DeAngelo* Anna Harvey†

February 22, 2018

*Associate Professor of Economics, Claremont Graduate University, email: gregory.deangelo@gmail.com
†Director, Public Safety Lab; Professor of Politics, New York University, email: anna.harvey@nyu.edu

The following projects are sponsored by the Public Safety Lab at NYU. If you choose to work on one of these projects, we will ask you to sign a nondisclosure agreement that safeguards the confidentiality of the data. We (Professors DeAngelo and Harvey) will also expect to meet with you regularly, both in person and via Skype, and to receive regular updates on your progress.

While we provide project ideas below, we are also open to creative suggestions for student projects using these data. Please contact Professor Harvey at anna.harvey@nyu.edu if you are interested in working with us.

1 Identifying Human Trafficking

Identifying human trafficking victims is very difficult for law enforcement agencies. Trafficked young women are held in isolation by their traffickers, coerced by violence, drugs, or both from attempting to flee or otherwise communicate with the outside world. They are also frequently moved by their traffickers to avoid detection. To solicit clients, traffickers place advertisements for their victims services on online message boards known to those who frequent commercial sex providers. Clients who review commercial sex providers may also post online reviews of providers who are actually trafficking victims.

Law enforcement agencies know that, among the more than 10 million commercial sex advertisements per year placed on online message boards, and among the hundreds of thousands of annual online provider reviews, there exist ads placed by traffickers and reviews of trafficking victims. But they do not know how to identify these ads and reviews.

DARPA's Memex project made initial steps toward the identification of human trafficking victims from online content. But prediction algorithms that can successfully identify human trafficking victims from online content have not yet been generated with a high degree of precision. The Public Safety Lab is carrying out an innovative project to build a prediction model that can accurately identify victims of human trafficking.

Our team has already harvested almost one billion online sex advertisements and approximately 2 million online provider reviews, and merged this content with a corpus of 25,000 true positive, 1,000 false positive, and 1,000 true negative phone numbers and email addresses sourced from law enforcement investigations of human trafficking. A small amount of social media data has been gathered, but more scraping of this content could be done.

Our current task is to extract content from the online ads and reviews that can eventually be used to generate a comprehensive prediction model for human trafficking. We welcome students interested in applying machine learning tools to these extractions.

Student project ideas:

- Use machine learning models to extract information from already harvested online content

that may be obfuscated, contained in images, etc.

- Use machine learning models to extract information from newly harvested online content, for example from social media platforms, that could be informative about human trafficking

In working on this project, students will learn how to use ML models to extract difficult content from various data sources.

Student skills required/desired:

Proficient in Python, Java, or similar language. Flexible in tool choices, familiar with big data architectures and standards. Familiarity with text search, extraction, analysis, and search solutions. Access to sufficient computing resources to work with very large data files required; experience working with large data files desired. Previous experience with Elasticsearch and Spark desirable.

2 Criminal Defendant Equity in New York State Courts

Understanding whether criminal defendants are treated equally in state and local courts, conditional on offense and other legally relevant criteria, is a policy question of critical importance. Yet it is a question typically confounded both by access to data and by problems of causal inference. Criminal case data do not typically include demographic information about criminal defendants, including race of defendant. Further, a considerable amount of legally relevant information is stored in the text of opinions, but is not accessible without sophisticated text analysis tools. Even if defendant demographic and legally relevant data were available, defendant demographic characteristics (which should not influence criminal justice decisions) are often closely associated with offense-related characteristics (which presumably should affect demographic characteristics). In this project, the Public Safety Lab is investigating the impact of criminal defendant demographic characteristics on trial and appellate court outcomes, conditional on legally relevant information.

We have scraped the text of every NYS appellate court opinion issued between 2003 and 2017; we have approximately 25,000 criminal appellate opinions in our data. We have also collected some and are in the process of collecting additional content external to the opinion texts.

Student project ideas:

- Use machine learning models to extract content from the text of judicial opinions that may be relevant to predicting the probability of reversal
- Build a prediction model to match defendant information extracted from opinion text to inmate demographic and criminal history information extracted from New York States Department of Corrections online lookup platform
- Build a prediction model to match appellate judge last names with appointment and election records extracted from online sources
- Build a prediction model to match trial judge last names with election records extracted from online sources

- Build a prediction model to match district attorney last names with election records extracted from online sources

- Using content extracted from opinion text, with or without additional content extracted from other sources, develop a prediction model of the probability of reversal, including heterogeneous effects conditional on defendant demographics

- Using content extracted from opinion text, along with additional content extracted from other sources about appellate judges, trial judges, and/or district attorneys, develop a prediction model of the probability of appellate reversal, including heterogeneous effects conditional on defendant demographics, using causal identification strategies based on appellate judge reappointment cycles, trial judge reelection cycles, district attorney reelection cycles, judge and district attorney pay increases, and/or campaign finance expenditures in districts that overlap trial judge and/or district attorney election districts.

In working on this project, students will learn a number of useful skills, including an understanding of criminal courts and their supporting institutions; how to apply ML models in the context of text extractions, particularly in the increasingly growing field of text extractions from legal documents; how to apply ML models in the context of matching and aligning records sourced from disparate venues; how to use ML models in the context of predicting outcomes in legal cases; and how causal identification techniques can be utilized when employing machine learning tools.

Student skills required/desired:

Proficient in Python, Java, or similar language. Flexible in tool choices, familiar with big data architectures and standards. Familiar with text search, extraction, analysis, and search solutions. Access to sufficient computing resources to work with large data files required; experience working with large data files desired. Previous experience with Elasticsearch and Spark desirable. Student with interest in data science, statistics, and data driven solutions desired. Statistical software (R, SAS, Stata) skills desired.