THESIS PROJECT PORTFOLIO

ACCURACY OF MACHINE LEARNING ALGORITHMS IN PREDICTING COLLEGE BASKETBALL GAMES

(Technical Report)

Predictions or Validation of Assumptions: Predictive Policing

(STS Research Paper)

An Undergraduate Thesis

Presented to the Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Fulfillment of the Requirements for the Degree
Bachelor of Science, School of Engineering

John Jordan Kim
Spring, 2022
Department of Computer Science

TABLE OF CONTENTS

SOCIOTECHNICAL SYNTHESIS

ACCURACY OF MACHINE LEARNING ALGORITHMS IN PREDICTING COLLEGE BASKETBALL GAMES
with Sindhura Mente
Technical advisor: Haiying Shen, Department of Computer Science

PREDICTIONS OF VALIDATION OF ASUSMPTIONS: PREDICTIVE POLICING
STS advisor: Kent Wayland, Department of Engineering and Society

PROSPECTUS
Technical Advisor: Haiying Shen, Department of Computer Science
STS advisor: Peter Norton, Department of Engineering and Society

The general research problem that was set out was, how can groups ethically use data to guide their decisions? The ethical conflict at play with data-driven decisions lies in defining the boundary between data extraction and privacy and the application of the data. Data-driven decisions have forms ranging from the use of targeted advertisements to sports analytics, and law enforcement using data for predictive policing. The scale for data extraction for commercial advertising is enormous as Google's website data service is present in 72.6% of the top 75,000 websites. The reason for this complex online architecture is the interest of the online technology companies in making money. Google made $133 billion USD from online advertisements alone in 2019. Using data to make predictions on people's online consumer tastes doesn't seem intrusive but making forecasts on people's disposition to crime, in the form of predictive policing, is an emerging front that has a greater impact on lives. It's calculated that law enforcement will spend $18.1 billion dollars on software for predictive policing. Using data to create predictive models raises the concern of where on the spectrum we allow data to make profiles on individuals and when the benefits of the practice outweigh the loss of privacy.

The technical aspect of the portfolio was concentrated on analyzing college basketball games to evaluate the effectiveness of different machine learning algorithms in predicting the outcomes of games. The motivation of the project lay in the immense popularity of the sport itself coupled with the amount of attention given to the field of sports betting. To execute the project data was scraped from basketball-reference and kenpom on all 358 NCAA Division 1 basketball teams from the 2021-2022 season. From there the data was fed into a pipeline to clean and prepare the data to be run and evaluated using four different models. The models were Linear Regression, Logistic Regression, Random Forest, and Decision trees. Using small samples, it was determined that Logistic Regression had the best predictive potential given the error was the smallest which was 0.59. Simulations of the NCAA Men's Basketball Tournament have been performed and it appears that the model hit a ceiling like other researchers of around 70%. The main conclusion that can be drawn from this is that there seems to be a limit to the extent that basketball statistics can give insight into who will be the winner. This cap on the accuracy may have something to do with the sporadic nature of sports in general and might demonstrate that the level of accuracy that numbers can give can't be surpassed.

The STS paper was focused on the topic of predictive policing and the evolving environment of actors that the practice lives in. The main research problem centered around the discussion of the trade-off of privacy and supporting predictive policing. A considerable marketed benefit of predictive policing was the effectiveness the practice had on reducing crime. This benefit was evaluated, and a major claim made was that looking over academic literature there seemed to be no consensus if predictive policing had a significant effect on crime. For the complications with predictive policing excerpts from legal professionals were used as evidence to make the claim that the legal bar of "reasonable suspicion" was being challenged and in turn, the fourth amendment was also inhibited. The use of studies of racial inequalities within the United States was used to support the claim that the controversial history of policing may impact the behavior of predictive policing. Analyzing the invasiveness of reverse search warrants constructed the claim that predictive policing was a hindrance to online privacy. In the conclusion, the final claim was made that if the principal benefit of crime being deterred is not even fully supported then the concerns regarding the fourth amendment, privacy, and reinforcing inequalities make the practice of predictive policing an unnecessary and may be exacerbating other problems.

At the completion of the technical and STS projects, all initial goals were met. Throughout the process of studying Machine learning on basketball datasets and looking at the effects and virtues of machine learning on criminal data a greater understanding of where data can succeed and fail to teach people about a topic was learned. For further research, in terms of the technical section, looking at the individual player performance, specifically the starters, on NCAA basketball teams may increase accuracy. For the STS section, searching for more literature on the legal boundaries predictive policing could run into may be fruitful.