A DEEP DIVE INTO GEORGIA'S RECIDIVISM

TEAM 3

Abhishek Anumalla

Nivedita J

Madhuri Muppa

Freny Patel

Lakshman Kushal Bogi

AIT614 Big Data Essentials

Dr. Liao

George Mason University

November 27, 2023

Abstract

This research endeavors to delve into the complexities of recidivism within the state of Georgia, adopting a multifaceted approach that encompasses data cleaning, exploratory data analysis (EDA), and predictive modeling. The primary goal is to understand the various factors influencing an individual's likelihood of reoffending within three years of their release. By examining demographic, behavioral, and systemic factors, the research aims to provide localized insights. Visualizations and machine learning models reveal patterns and correlations, offering a foundation for targeted interventions to reduce recidivism rates and enhance the overall effectiveness of rehabilitation strategies.

The study recognizes recidivism as a persistent issue within the criminal justice system and emphasizes its impact on individuals and society. By focusing specifically on Georgia, the research aims to provide localized insights, steering away from generalized conclusions drawn from broader datasets. The research outlines specific objectives, including an examination of demographic factors such as gender, race, and age in relation to recidivism rates. It further explores the correlation between prior criminal history and reoffending, investigates the role of behavioral factors, and assesses the effectiveness of supervision levels. The study also aims to identify patterns and trends in recidivism over specified time frames and analyze systemic and societal factors.

Exploratory data analysis serves as a critical component in unraveling the complexities of reoffending. Visualizations, ranging from bar to line charts, illuminate the relationships between demographic factors, behavioral traits, and recidivism. The analysis dives into the details of age, gender, ethnicity, education, and various aspects of criminal history, aiming to discern patterns

and correlations that can inform targeted interventions. This research employs Databricks DBFS, a NoSQL database, for efficient data storage and management. PySpark is utilized for data preprocessing and transformation, R is used for data visualizations, while Spark MLlib facilitates the development of predictive models.

The predictive modeling phase employs a spectrum of machine learning algorithms, including logistic regression, random forests, support vector machines, and gradient boosting. These models undergo refinement through hyperparameter tuning and cross-validation, ensuring robust predictions. The research extends beyond prediction, incorporating chi-square testing for feature selection. This step identifies the most influential features impacting recidivism, contributing to a nuanced understanding of the driving factors.

The outcomes of this research offer a comprehensive perspective on recidivism, embracing the multifaceted nature of demographics, behaviors, and systemic influences. The findings contribute valuable insights to evidence-based decision-making within the criminal justice system. The collaborative fusion of traditional statistical methods and advanced machine learning underscores the need for a holistic approach to address this societal challenge.

In conclusion, this project represents a significant stride in comprehending and addressing recidivism in Georgia. The combination of rigorous data processing, exploratory analysis, and advanced predictive modeling provides a nuanced understanding of the intricate interplay of factors influencing reoffending. The insights garnered from this research have the potential to guide policy decisions, enhance intervention strategies, and foster a more effective criminal justice system.

Team 3

Keywords: Recidivism, Georgia, Multifaceted approach, Data cleaning, Exploratory data analysis (EDA), Predictive modeling, Demographic factors, Visualizations, Machine learning models, Correlations, Criminal justice system, Databricks DBFS, NoSQL database, PySpark, R, Spark MLlib, Logistic regression, Random forests, Support vector machines, Gradient boosting, Hyperparameter tuning, Cross-validation, Chi-square testing, Feature selection

References

[1] Dr.Liao Getting Started with MLlib.html Blackboard, https://mymasonportal.gmu.edu/bbcswebdav/pid-18409737-dt-content-rid-292661466_1/xid-292661466_1

[2] G. Department, “NIJ’s Recidivism Challenge Test Dataset1,” Usdoj.gov, Apr. 23, 2021. https://data.ojp.usdoj.gov/Corrections/NIJ-s-Recidivism-Challenge-Test-Dataset1/d7q7-hb2x/about_data (accessed Dec. 27, 2024).

[3] MLlib | Apache Spark. (2019). Apache.org. https://spark.apache.org/mllib/