LISBON DATA SCIENCE ACADEMY

Fairness in Recidivism Prediction: A Data-Driven Approach

A Comprehensive Analysis and Model Development Approach for Fairer Criminal Justice Decision-Making

Prepared for:
Florida Department of Corrections

Prepared by:
João Pinto
Data Scientist
Awkward Problems Solutions™

22 of May 2024

Table Of Contents

1. Project Definition

1.1 Introduction

The Florida Department of Corrections (FDOC) has engaged our consulting firm, Awkward Problem Solutions™, to address concerns regarding the fairness of their current recidivism prediction algorithm. This algorithm, previously developed by another company, is suspected of bias against certain ethnic and socio-demographic groups. Given the high stakes involved in judicial decisions and the significant societal implications, it is imperative that the algorithm is fair and unbiased. Our task is to investigate the existing algorithm for any signs of bias, propose corrective measures if necessary, and develop a new predictive model to be deployed via an API. This model will support judges and probation officers in assessing the likelihood of recidivism among criminal defendants, ensuring equitable treatment across all demographics.

1.2 Objectives

The objectives of this project are multifaceted, aiming to address the core challenges identified by the Florida Department of Corrections and ensure the development of a fair and effective predictive model for recidivism:

1. **Bias Detection and Mitigation:** The primary objective is to meticulously examine the current recidivism prediction algorithm to detect any inherent biases, particularly against specific ethnic or socio-demographic groups. Through rigorous analysis and fairness assessments, we aim to identify and mitigate any biases present in the existing algorithm.

2. **Development of a Fair Predictive Model:** Building upon the findings of bias detection, our next objective is to develop a new predictive model for recidivism that prioritizes fairness and equity across all demographic groups. We will strive to create a model that not only accurately predicts recidivism but also ensures that predictions are unbiased and equitable for all individuals, regardless of race, ethnicity, or other socio-demographic factors.

3. **Deployment of a Reliable API Endpoint**: In addition to developing the predictive model, we aim to deploy it as a RESTful API endpoint that seamlessly integrates with the existing systems of the Florida Department of Corrections. This API will provide real-time predictions of recidivism risk, enabling judges and probation officers to make informed decisions based on the latest data and insights, ensuring the reliability, scalability, and security of the entire process.

1.3 Scope

This project will include several key tasks. We will begin by cleaning and preprocessing the provided dataset to ensure it is suitable for analysis and modeling. Next, we will conduct exploratory data analysis to identify patterns, trends, and potential biases in the data, focusing on personal data such as name, sex, age, and race, case data like jail dates, charge degree, and the number of prior crimes, data on previous arrests, and recidivism scores. We will then develop fairness metrics to assess the current algorithm's performance across different ethnic and socio-demographic groups. After evaluating various machine learning algorithms, we will develop a new predictive model to ensure it is unbiased and fair.Finally, we will deploy this model as a REST API, enabling real-time predictions, ensuring

compliance with security standards and recieval of the recieval of the remaining training data. After receiving this data, the model will be retrained and adjusted as necessary.
The project will exclude external data collection, real-time data streaming and post-deployment monitoring and maintenance of the API.

2. Project details

2.1 Methodology

Our project methodology begins with a thorough data preprocessing and exploratory data analysis (EDA) phase. We will start by meticulously cleaning and preprocessing the provided dataset. This involves handling missing values, outliers, and ensuring data consistency while anonymizing personal information to uphold data privacy standards. Subsequently, we will conduct thorough additional EDA to discern underlying patterns, trends, and potential biases within the data. Specifically, we will scrutinize variables such as personal data (e.g., name, sex, age, race), case data (e.g., jail dates, charge degree, number of prior crimes), and recidivism scores. Our focus will then shift towards detecting and addressing biases inherent in the current recidivism prediction algorithm. We will employ a range of fairness metrics, including disparate impact ratio, equal opportunity difference, and average odds difference, to evaluate the algorithm's performance across different demographic groups. To mitigate identified biases, we will apply techniques such as reweighting, adversarial debiasing, and fairness-constrained optimization, ensuring fairness in the predictive model.

With a comprehensive understanding of the data and fairness considerations, we will proceed to develop a new predictive model for recidivism. Our primary objective is to construct a robust predictive model for recidivism, tailored specifically to address the classification problem at hand—predicting whether an individual will reoffend or not. Our plan follows a systematic approach: We will start by selecting and engineering features relevant to predicting recidivism, aiming to capture meaningful predictors while minimizing noise. Afterward, we'll evaluate multiple classification algorithms to determine the most suitable model, prioritizing performance in binary classification tasks. Once we've selected the models, we'll proceed to train and validate them using techniques such as k-fold cross-validation, optimizing hyperparameters to maximize predictive accuracy while avoiding overfitting. Throughout this process, fairness considerations will be integrated to ensure equitable predictions across different demographic groups. Furthermore, we'll emphasize model interpretability to enhance transparency and facilitate stakeholder understanding of prediction outcomes. Finally, we will choose the final model based on its predictive accuracy, fairness, interpretability, and scalability, aligning with the project's objectives and ethical considerations.

Once the predictive model is developed and validated, we will deploy it as a RESTful API endpoint. Leveraging Python-based web frameworks (FLASK), we will develop a robust and scalable API that seamlessly integrates with the Florida Department of Corrections' internal systems and guarantees the reception of the remaining training data. Upon receiving the remaining data, the model will be retrained according to the new information and redeployed to ensure maximum accuracy. Continuous monitoring and testing will be conducted to ensure the reliability and performance of the API endpoint in real-world scenarios.

Finally, throughout the project, meticulous documentation of our methodology, analysis, model development process, and deployment procedures will be maintained. This documentation will serve as a transparent record of our efforts and facilitate knowledge transfer to stakeholders within the Florida Department of Corrections. Additionally, a detailed report summarizing our findings, methodologies, and recommendations will be prepared. This report will provide insights into the fairness of the predictive model and its implications for decision-making within the criminal justice system.

2.2 Timeline

Our project timeline is structured to align with the given deadlines, ensuring systematic progress and timely completion of all deliverables. Here's a detailed breakdown of the timeline:

6 - 12 May: Data Preprocessing and Exploratory Data Analysis (EDA) - Data Cleaning, EDA, Clarifications

13 - 15 May: Clarifications from Employer

16 - 22 May: Proposal Preparation

22 - 26 May: Proposal Review and Feedback

26 May - 16 June: Model Development and Initial API Preparation - Feature Selection and Engineering, Model Selection, Training and Validation, Fairness Integration**,** Model Interpretability, API Development and Initial Testing

9 - 16 June: Trial Round and Final Adjustments - Conduct a trial round for testing the API.

16 June: API and Report Delivery Deadline - Submit the API and the report.

17 - 23 June: Employer Testing and Review - API Testing and Report Review

24 June - 7 July: Address Feedback and Retrain Model - Feedback Implementation, Model Retraining and Redeployment

7 - 14 July: Final Review

3. Interpretation

3.1 Expected Outcomes

By the end of this project, we anticipate achieving several significant outcomes. Firstly, we will deliver a comprehensive analysis of the current recidivism prediction algorithm utilized by the Florida Department of Corrections. This analysis will identify any existing biases or fairness issues, focusing on how the algorithm performs across different demographic groups. We will use fairness metrics such as disparate impact ratio, equal opportunity difference, and average odds difference to provide a detailed assessment.

Secondly, we will develop a new predictive model specifically designed to address and mitigate these biases, ensuring that predictions are fair and accurate for all demographic groups. This model will be deployed as a RESTful API, fully integrated with the Florida Department of Corrections' internal systems to enable seamless usage by judges and probation officers. Additionally, upon receiving the remaining training data, we will retrain the model to enhance its predictive capabilities further.

Finally, we will compile a detailed final report that documents our entire process—from data preprocessing and bias detection to model development and deployment. This report will include our findings, methodologies, and recommendations for ongoing improvements, providing valuable insights for stakeholders to support fair and effective decision-making in the criminal justice system.

3.2 Evaluation

The success of this project will be evaluated based on several criteria. Firstly, the accuracy of the new predictive model will be assessed using standard classification metrics such as accuracy, precision, recall, and F1-score. Additionally, the fairness of the model will be evaluated using the aforementioned fairness metrics, ensuring that the model does not exhibit significant bias against any demographic group. Lastly, the comprehensiveness and clarity of the final report will be assessed to ensure that it provides valuable insights and actionable recommendations.

3.3 Conclusion

In conclusion, this project aims to tackle the critical issue of fairness in recidivism prediction within the Florida Department of Corrections. By meticulously cleaning and preprocessing the data, followed by a comprehensive exploratory data analysis, we will uncover any inherent biases and trends in the current algorithm. Our detailed analysis will provide a clear picture of how the existing model performs across various demographic groups, using robust fairness metrics to ensure thorough evaluation.

The development of a new predictive model is central to our efforts. This model will be designed not only to improve predictive accuracy but also to address and mitigate identified biases. By employing machine learning techniques and integrating fairness considerations, we will strive to create a model that makes equitable predictions for all individuals, regardless of their background. The model's deployment as a RESTful API will facilitate its integration into the Florida Department of Corrections' systems, ensuring it is a practical tool for judges and probation officers in assessing recidivism risk.

Upon receiving additional training data, we will retrain the model to further refine its accuracy and reliability. Continuous monitoring and testing of the API will be conducted to ensure its performance remains robust in real-world scenarios.

The final deliverable will be a comprehensive report documenting our entire process, from initial data preprocessing to the deployment of the predictive model. This report will include detailed findings, methodologies, and recommendations, providing a transparent and thorough account of our work. It will serve as a valuable resource for stakeholders, offering insights into the fairness of the predictive model and its implications for decision-making within the criminal justice system. By addressing these critical issues, we aim to contribute to more equitable and informed outcomes in the assessment of recidivism risk, ultimately supporting the broader goals of justice and fairness.