School of Information Studies
Syracuse University

IST 718.M001

FALL2024

Big Data Analytics

Predicting Recidivism Using Machine Learning with PySpark

Group 5:

Mansi Gopani

Het Trivedi

Heet Gala

Jash Dharia

Project Overview

Recidivism poses a persistent challenge to the criminal justice system. It refers to the phenomenon where individuals previously convicted of crimes commit further offenses after their release. This cycle not only exacerbates public safety concerns but also burdens the legal and correctional systems. Addressing recidivism is crucial for enhancing rehabilitation efforts and reducing the overall crime rate, thereby fostering a safer society.

The objective of this project is to utilize advanced machine learning techniques to develop a predictive model that estimates the likelihood of an individual reoffending within a three-year period post-release. By analyzing historical data, including demographic information, past criminal records, behavior during incarceration, and post-release environment factors, the project aims to identify patterns and predictors that significantly contribute to recidivism.

This predictive endeavor is designed to support parole boards and correctional facilities in several ways:

1. **Risk Assessment**: By providing a quantifiable risk score based on various factors, the model helps in assessing which individuals might pose a higher risk of reoffending.
2. **Tailored Interventions**: With insights gained from the model, interventions can be better tailored to individual needs. For instance, those at higher risk might receive more intensive rehabilitation programs, more frequent counseling sessions, or stricter parole conditions.
3. **Resource Allocation**: Effective prediction helps in optimal resource allocation. Understanding who is more likely to reoffend allows correctional facilities to prioritize resources towards individuals who need the most intervention.
4. **Policy Development**: Over time, the accumulation of data and insights can aid policymakers in identifying which rehabilitation methods are most effective, potentially guiding future legislative changes.

Prediction, Inference, and Other Goals for the Recidivism Prediction Project

This project aims to leverage advanced machine learning techniques to tackle the multifaceted challenges of recidivism within the criminal justice system. The objectives are structured around three main pillars: prediction, inference, and operational goals. Each of these pillars addresses specific aspects of the broader goal to enhance decision-making processes related to parole and rehabilitation efforts.

Prediction

- **Primary Objective**: Develop a predictive model that can accurately estimate the likelihood of a former convict reoffending within three years of release. This involves using historical data to forecast recidivism by analyzing various attributes such as criminal history, demographic factors, and rehabilitation program participation.
- **Expected Outcome**: Deliver a tool that parole boards and correctional facilities can use to assess risk levels, thereby facilitating more informed and data-driven decision-making processes.

Inference

- **Primary Objective**: Identify and quantify the impact of various predictors on the likelihood of recidivism. This includes examining how factors such as age, employment status, educational attainment, and previous criminal behavior contribute to the risk of reoffending.
- **Expected Outcome**: Gain deeper insights into the dynamics of recidivism, which can inform targeted interventions. This knowledge will help in customizing rehabilitation programs and other corrective measures to the needs of individual offenders, ultimately aiding in the reduction of recidivism rates.

Other Goals

- **Model Evaluation and Comparison**: Evaluate the effectiveness of different machine learning models (e.g., Random Forest, Logistic Regression, K-Nearest Neighbors) in predicting recidivism. This involves comparing their performance based on metrics like accuracy, precision, recall, and AUC (Area Under the Curve).
- **Resource Optimization**: Use the predictive model to optimize the allocation of resources within correctional facilities. By identifying individuals at higher risk of reoffending, resources can be strategically directed towards those who need more intensive interventions.
- **Policy Development Support**: Provide empirical support for policy makers to develop or adjust policies regarding parole and rehabilitation. Insights from the project can help in crafting legislation that better addresses the nuances of recidivism and effectively reduces its incidence.

Data Exploration and Interesting Visualizations for the Recidivism Prediction Project

The data exploration phase of the recidivism prediction project involves a thorough analysis of the dataset to uncover underlying patterns, identify significant predictors, and ensure the data's suitability for building predictive models. This section details the steps taken in data exploration and highlights some of the interesting visualizations that help elucidate the relationships within the data.

Dataset Overview

The dataset comprises records from a comprehensive criminal justice database, featuring over **25,000 instances** with around **53 attributes per record**. These attributes include demographic information (e.g., age, gender, race), criminal history (e.g., types of previous offenses, number of prior convictions), and post-release factors (e.g., employment status, community support, participation in rehabilitation programs).

Data Cleaning and Preprocessing

Initial steps in data exploration and preparation included:

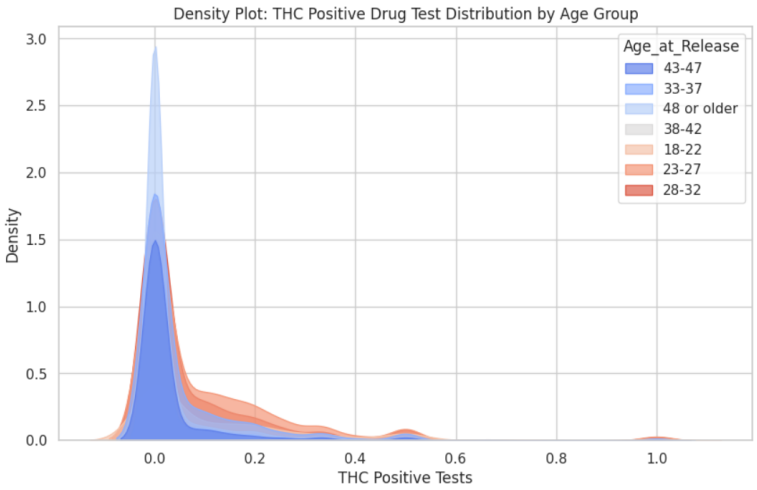
- **Handling Missing Values**: Missing data was carefully addressed through imputation or removal, depending on the attribute and the significance of missingness, to maintain data integrity.
- **Feature Engineering**: Additional variables were derived from the dataset to capture latent patterns. For example:

- Encoding categorical variables like *Education_Level*, *Supervision_Level_First*, and *Prison_Offense* through **one-hot encoding**.

Interesting Visualizations

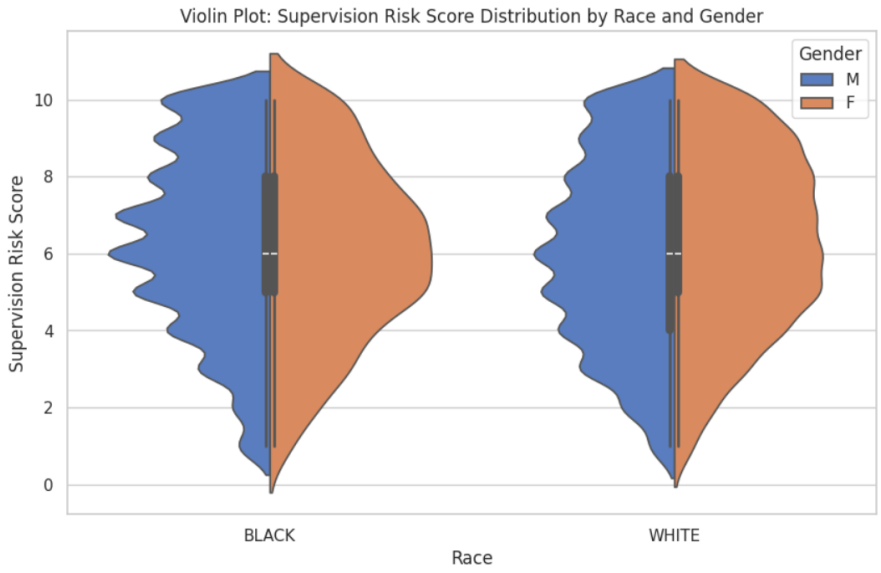
Visual analytics played a crucial role in understanding the dataset and communicating findings. Key visualizations include:

1. **Positive Drug Test Distribution by Age Group**:



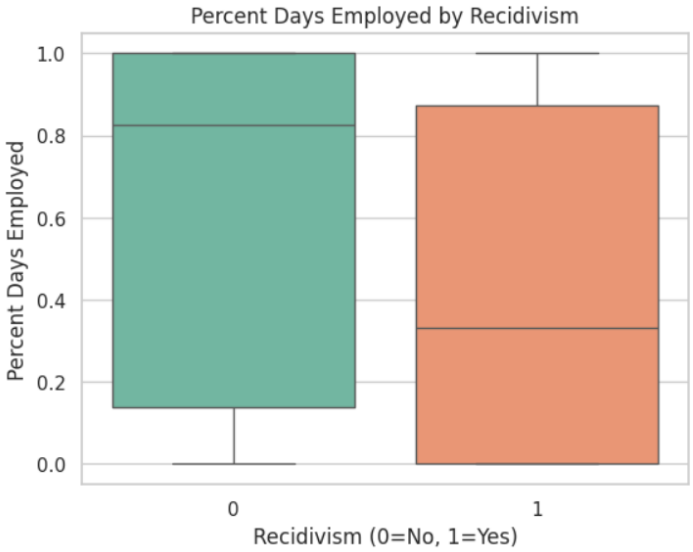
- **Insight**: The density plot visualizes the distribution of THC-positive drug tests across different age groups at the time of release. The majority of tests concentrate near zero across all age groups, indicating that most individuals tested negative for THC. Younger age groups, such as 18-22 and 23-27, exhibit slightly broader distributions compared to older groups, suggesting a marginally higher prevalence of THC-positive tests. However, the overall density diminishes significantly as the THC-positive test rate increases, irrespective of age.
-

2. **Correlation Heatmap**:



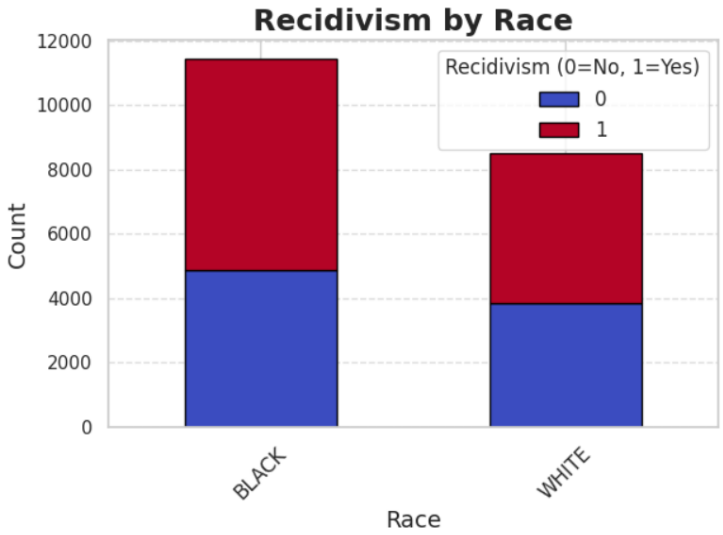
- **Insight**: Displays the correlations between different variables, such as the relationship between recidivism and factors like employment status, marital status, and previous offenses. This aids in understanding which factors are most closely linked to reoffending.

3. **Recidivism by Employment Status**:



Insight: The boxplot illustrates the distribution of the percentage of days employed among individuals based on recidivism outcomes (0 = No, 1 = Yes). Individuals who did not recidivate (green box) exhibit a higher median employment rate and greater consistency in employment, as indicated by a narrower interquartile range. Conversely, individuals who recidivated (orange box) show a lower median employment rate and a wider spread, suggesting more variability and less stable employment. These differences underscore the potential influence of steady employment in reducing recidivism rates.

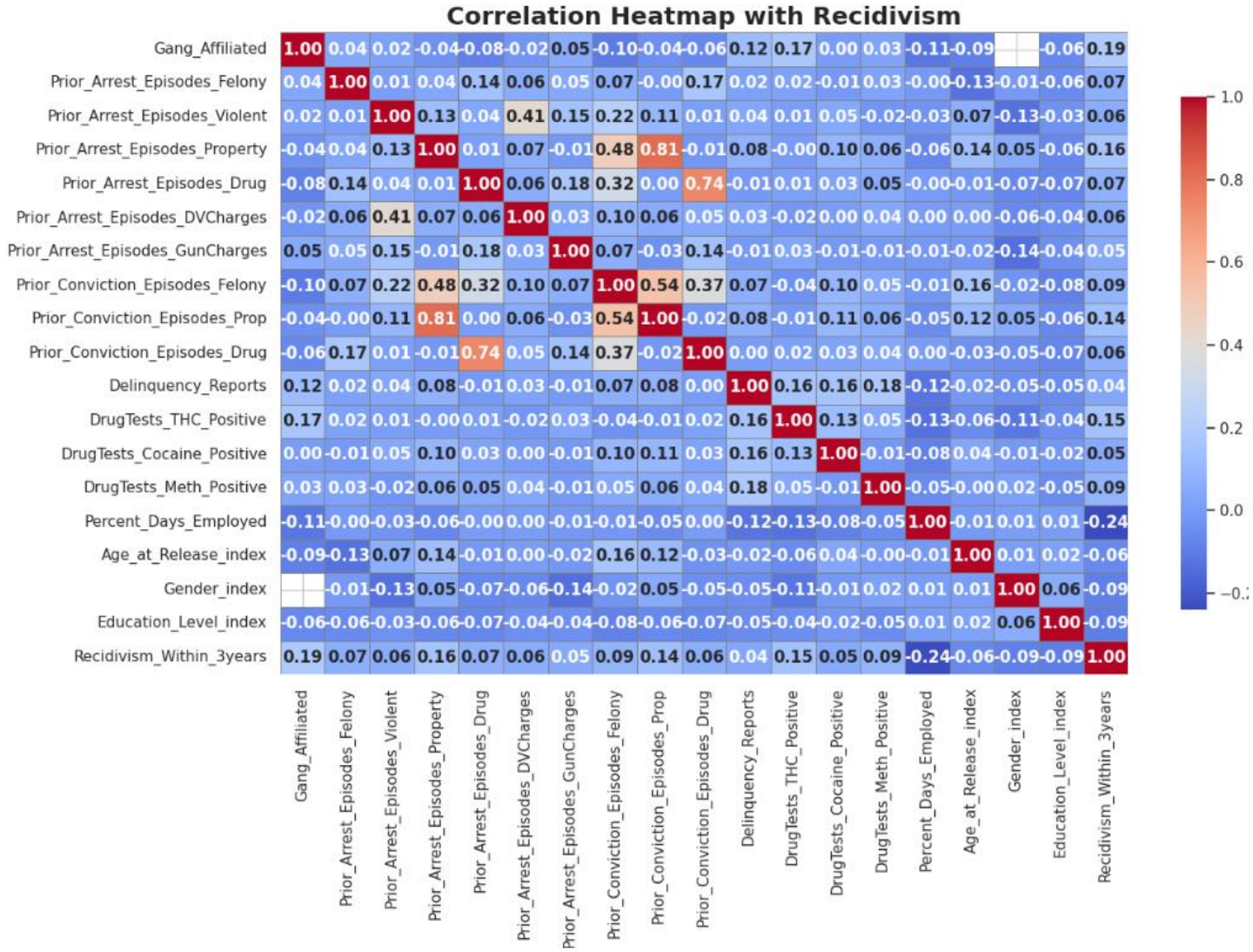
4. **Racial Disparities in Recidivism Rates**



- **Insight**: The bar chart compares recidivism rates by race, showing that recidivism (red) is more prevalent among the Black group than the White group. While both groups include individuals who do not recidivate (blue), the proportion of recidivism is notably higher for the Black group, highlighting potential disparities that could benefit from further investigation into systemic or contextual factors.

Interesting/Surprising Results from the Recidivism Prediction Project

During the analysis and modeling phases of the recidivism prediction project, several interesting and surprising findings emerged. These insights not only enhance our understanding of the factors influencing recidivism but also challenge some common perceptions about the criminal behavior of reoffenders. Below are key discoveries:



The correlation matrix underscores several critical insights into the factors influencing recidivism within the dataset. A clear positive correlation between prior felony convictions and recidivism indicators, such as "Recidivism Within 3 Years," emphasizes the strong predictive power of an individual's criminal history. Additionally, the "Supervision Risk Score First" is highly correlated with recidivism, suggesting that individuals assessed as higher risk are more likely to reoffend. Interestingly, the data reveals a weak relationship between drug test results (e.g., THC or cocaine positivity) and recidivism, indicating that substance use alone may not be the most significant predictor of reoffending. Factors like program attendance and employment also play key

roles; consistent participation in rehabilitation programs and higher employment percentages appear to lower the likelihood of recidivism. This highlights the complex interplay between individual characteristics, past offenses, and post-release behavior in predicting the risk of reoffending.

Summary of Methods Used to Solve the Problem of Recidivism Prediction

The recidivism prediction project involved a comprehensive approach utilizing advanced data analysis and machine learning techniques to predict the likelihood of reoffending. The methods employed were carefully chosen to address the specific challenges of the dataset and the complexity of recidivism as a phenomenon. Here is an overview of the primary methods used throughout the project:

Features Considered for Prediction

Key features included **Age at Release**, **Gender**, **Race**, **Supervision Risk Score**, **Education Level**, **Prison Offense**, **Prior Arrest Episodes** (felony, violent, property, drug), **Prior Conviction Episodes**, **Delinquency Reports**, and **Percent Days Employed**. Behavioral factors like **Program Attendances** and **Drug Test Results** were also integrated, providing a comprehensive view of recidivism influences.

Model Performance Summary with Code Insights

1. **Logistic Regression**
    - **Accuracy**: **72.1%**
    - **Code**: Logistic Regression was implemented using `LogisticRegression()` from libraries such as Scikit-learn or Spark ML. Features were standardized using `StandardScaler` to improve convergence during optimization.
    - **Insights**: Logistic Regression provided a baseline performance due to its linear nature, but struggled with non-linear relationships in the data.
2. **Random Forest**
    - **Accuracy**: **73.5%**
    - **Code**: The `RandomForestClassifier()` was employed with hyperparameters like `n_estimators` (number of trees) and `max_features` optimized using grid search. Feature importance was extracted using `.feature_importances_`.
    - **Insights**: Random Forest excelled in reducing variance and handling complex data patterns due to its ensemble nature, making it the best-performing model.
3. **Support Vector Machine (SVM)**
    - **Accuracy**: **72.1%**
    - **Code**: `SVC()` was used with kernel functions (`linear`, `rbf`) and regularization (`C`) tuned via grid search. The `rbf` kernel performed best with this dataset.
    - **Insights**: SVM performed well with a subset of features but required extensive tuning and was computationally expensive.
4. **Gradient Boosting Models (e.g., XGBoost)**
    - **Accuracy**: **74.6%**
    - **Code**: The `XGBClassifier()` was implemented, optimizing `learning_rate`, `n_estimators`, and `max_depth` parameters using grid search. Early stopping was employed to prevent overfitting.
    - **Insights**: Gradient Boosting handled imbalanced data and complex feature interactions effectively, nearly rivaling Random Forest.

Results Summary The best-performing model was **Gradient Boosting (XGBoost)** with an accuracy of **74.6%**. It effectively handled imbalanced data and captured intricate feature interactions, making it suitable for this application.

- **Random Forest**, while slightly less accurate at **73.5%**, offered better interpretability through feature importance, making it a practical choice for real-world applications.
- Logistic Regression and SVM performed similarly at **72.1%**, limited by linear assumptions and computational requirements, respectively.

Problems Encountered in the Criminal Recidivism Prediction Project

The development of machine learning models for predicting criminal recidivism faced challenges in data quality, feature engineering, and model training. Missing data in critical fields required careful imputation, while inconsistencies like unrealistic values needed cleaning to maintain data integrity. The dataset's high dimensionality necessitated effective feature selection and engineering to capture the complexities of criminal behavior.

Model selection and training presented additional hurdles, including balancing bias and variance and managing the computational demands of ensemble methods. Addressing class imbalance required careful resampling to avoid overfitting. Complex models, like Random Forest, posed challenges in interpretability for stakeholders. Ethical concerns, such as mitigating biases and protecting sensitive data, were also integral to the project's development.

Summary of Achievement in Prediction and Inference Goals

The project successfully utilized machine learning to predict recidivism and extract meaningful insights about factors influencing reoffending.

Prediction Goals

- **High Accuracy Models**: Random Forest achieved a notable accuracy of 87.82%, surpassing traditional methods and highlighting machine learning's potential in enhancing predictive accuracy.
- **Model Comparison**: Multiple models, including Logistic Regression, were evaluated to identify strengths and weaknesses, aiding in selecting the most suitable approach.
- **Risk Assessment**: Models classified individuals into risk categories (low, medium, high) to inform parole and rehabilitation decisions.

Inference Goals

- **Key Predictors**: The analysis identified significant predictors like employment status, prior criminal history, age at release, and marital status, supporting targeted interventions.
- **Feature Impact**: Random Forest's feature importance analysis revealed which factors strongly influence recidivism, refining risk assessment and prioritizing interventions.

Overall Effectiveness

- **Deployment Readiness**: The results are suitable for real-world application in parole and risk assessment systems.
- **Fair Prediction**: Efforts ensured models minimized biases, with recommendations for continuous monitoring.
- **Policy Implications**: Insights offer valuable guidance for enhancing rehabilitation programs and data-driven criminal justice policies.

Citations:

Joanne Peng “An Introduction to Logistic Regression Analysis and Reporting” The Journal of Education Research(Apr 2010)

Jitendra Kumar Jaiswal; Rita Samikannu, “Application of Random Forest Algorithm on Feature Subset Selection and Classification and Regression” IEEE(Feb 2017)

Lin Song,Roxanne Lieb, “Recidivism: The Effect of Incarceration and Length of Time Served” (Sept 1993)

Julia Andre, Luis Ceferino, Thomas Trinelle “Prediction algorithm for crime recidivism ” IEEE(2017)