Analysis

Overview

This repository contains a comprehensive analysis of reincarceration data using Python. The analysis covers data preprocessing, feature engineering, exploratory data analysis (EDA), modeling using statistical and machine learning approaches, and time series forecasting. Follow the steps below to understand and execute the analysis.

Files

There is a set of three notebooks, each with its specific focus and analysis:

1. Incarcerated Days Notebook:

- This notebook is dedicated to the implementation and tuning of machine learning models and statistical analysis. The focus is on predicting and understanding the duration of incarceration for individuals. The models utilized here may include machine learning algorithms, and statistical techniques are applied to analyze and interpret patterns in the data.

2. Duration out of Jails & Incarceration Counts Notebook:

- In this notebook, statistical analysis takes center stage. It involves exploring and interpreting quantitative aspects related to incarceration, such as counts, demographic breakdowns, or other numerical metrics. Statistical methods and machine learning models are applied to extract insights from the dataset without a specific emphasis on time series analysis.

3. Re-incarceration Notebook:

- This notebook covers a broader spectrum by incorporating statistical analysis, machine learning models, and time series analysis. It addresses the specific aspect of re-incarceration, exploring patterns over time. Statistical methods are employed for quantitative insights, machine learning models for predictive analysis, and time series models for understanding temporal trends in re-incarceration occurrences.

Collectively, these three notebooks cover a broad spectrum of analyses, incorporating machine learning, statistical techniques, and time series methods to provide a comprehensive understanding of different aspects related to incarceration. Each notebook contributes a unique perspective to the overall analysis, making the combined results more robust and informative.

Dependencies

Make sure you have the following Python libraries installed:

- pandas

- numpy
- matplotlib
- seaborn
- sklearn
- statsmodels
- xgboost
- tensorflow

You can install these dependencies using the following command:

```
pip install pandas numpy matplotlib seaborn scikit-learn
statsmodels xgboost tensorflow
```

Usage

Follow the detailed steps in the notebook to execute the analysis. The notebook is structured as follows:

1. Data Loading and Preprocessing:
    - Load the reincarceration data from the CSV file (CensusJDI.csv) and preprocess it, handling missing values and data types.
2. Feature Engineering:
    - Create new features, calculate time gaps, and count reincarceration occurrences for each individual.
3. Exploratory Data Analysis (EDA):
    - Explore the data through visualizations and statistical summaries to gain insights into patterns and trends.
4. County-level Analysis:
    - Analyze reincarceration rates and trends at the county level, including calculations of rates and averages.
5. Principal Component Analysis (PCA):
    - Apply PCA to understand the relationships and importance of features in the dataset.
6. Linear Regression and Lasso Regression:
    - Perform linear regression and lasso regression to analyze the impact of features on reincarceration.
7. Mixed-Effects Modeling:
    - Apply mixed-effects models to account for random effects at the county level.
8. Likelihood Ratio Test:
    - Conduct a likelihood ratio test to assess the significance of the mixed-effects model.
9. XGBoost Model:
    - Train an XGBoost model to predict reincarceration and evaluate its performance.

10. Neural Network Model:
    - Implement a neural network model using TensorFlow for binary classification.
11. Time Series Forecasting:
    - Utilize SARIMA models to forecast reincarceration counts for each county.

Data

The analysis uses a CSV file named CensusJDI.csv. Ensure that this file is present in the same directory as the Jupyter Notebook. The data should include these columns:

- 'nameFull':'str',
- 'bookingNumber':'str',
- 'caseNumber':'str',
- 'inmateNumber':'str',
- 'housing1':'str',
- 'housing1a':'str',
- 'housing2':'str',
- 'housing3': 'str',
- 'nameLast':'str',
- 'nameFirst':'str',
- 'nameMiddle':'str',
- 'age': 'float',
- 'sex': 'category',
- 'race': 'category',
- 'eyeColor': 'category',
- 'hairColor': 'category',
- 'height': 'float',
- 'weight': 'float',
- 'bond': 'float',
- 'bail': 'float',
- 'topcharge': 'category',
- 'county':'category',
- 'Top Poliytical': 'category'

Results

The notebook generates various visualizations, statistical summaries, and model evaluations throughout the analysis. The final output includes visualizations of reincarceration trends, feature importance scores, and time series forecasts for each county.

Notes

- Some parts of the code may require modifications based on your specific dataset or analysis goals
- Ensure you have the required permissions to access and use the data.

Feel free to reach out if you have any questions or need further assistance. Happy