ORIE 5741 Final Project Proposal

Zongjie Yin(zy347), Chunyu Chen(cc2582)

March 19, 2023

Dataset Introductions

The data contains 54 columns in total with 25835 rows. It recorded each criminal's intrinsic data like gender, race, etc. Besides, it keeps track each criminal's crime activities data like prior crime type, prison offense type, risk score, and recidivate or not in the next three years that given by the Kaggle

Questions to explore

Main Question:

- Predict whether the criminal will recidivate in the next three years?

Subquestions:

- Whether the race or gender difference will affect the recidivism?
- Whether the prior crime type will affect the recidivism?
- Whether the risk score or risk level will affect the recidivism?
- Is there any distinct pattern across race, gender, or crime type?

Why the problem is important?

This problem is important because in the US, criminals are sentenced based on the Compas Score given by the Compas Algorithm, which sometimes produces biased or unreasonable scores that can lead to wrong decisions. By helping the government better understand which inmates are likely to reoffend, we can control the number of criminals in jail while improving the accuracy of recidivism predictions

Why the data set will allow me to answer the question?

The dataset is sufficiently large (25835 records) with comprehensive intrinsic (gender, race, etc.) and crime activity data and recidivism outcomes, making it a suitable resource to answer the research questions

Why it is worthwhile for you to work on?

It can let the government control the number of criminals in jail, at the same time, ensure the accuracy of prediction of recidivism

Why you think you are likely to succeed?

- We can make use of the SVM, logistics regression, Decision Tree, and Random Forest, even the neural network technique to do the binary response prediction (whether the criminal recidivate)
- We can also use PCA to reduce the dimension of the dataset and K-means clustering to identify distinct patterns
- Additionally, the dataset is large enough for us to perform train-test split and cross-validation for model evaluation

Reference

https://www.kaggle.com/datasets/uocoeeds/recidivism