ORIE 5741 Final Project Proposal

Zongjie Yin(zy347), Chunyu Chen(cc2582)

March 19, 2023

Dataset Introduction

The data contains 53 columns in total with 7215 rows. It recorded each criminal's intrinsic data like sex, age, etc. Besides, it keeps track each criminal's crime activities data like crime type, whether he/she recidivate in the next two years after jail out, and the Compas score that given by the Compas Algorithm.

Questions to explore

Main Question:

- Predict whether the criminal will recidivate in the future?

Subquestions:

- Whether the race, nationality, and skin color differences will affect the COMPAS SCORE, which may affect the prediction of recidivism?
- Whether the age differences will affect the Compas Score?
- Whether the sex differences will affect the Compas Score?
- Whether the parole affects the recidivation?
- Is there any data that are misleading or untrusty?

Why the problem is important?

In the US, the counts will give the criminal a sentence based on their experiment and the Compas Score given by the Compas Algorithm. However, it sometimes gives the unreasonable or bias score, which misleads the counts and let them make the wrong decision. We aim at helping government understand which inmates are likely to commit a subsequent crime, and which inmates are ready for parole.

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

We have 7215 records in total, which is enough even though it is not that "big." We have the intrinsic data (sex, age, etc.) of the suspicious criminal and the crime activities data, and whether they will recidivate in the future (next two year). So, we think the data set will allow me to answer the question.

Why it is worthwhile for you to work on?

It can let the government control the number of criminals in the jail, at the same time, assure the accuracy of prediction of recidivism and parole.

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 parole and whether the criminal recidivate);
- We can utilize the PCA to lower the dimension of the data set;
- The data set is quite complete and the number of data is big enough for us to do the train-test split.