**Project 3: Interpretability**

#627

Project 3

For Project 3, our primary focus lies in the interpretability of the classifier that you construct once the accuracy requirement is met.

Consider the following hypothetical scenario:

Airbnb hosts are primarily concerned with avoiding undesirable guests. Despite Airbnb recently making host reviews of guests public, there are certain limitations. Many reviews are absent, and more importantly, hosts are hesitant to openly criticize guests in public reviews, as this may deter potential guests from booking their accommodations.

Given that most guests have left reviews for their stays, Airbnb aims to create a classification model capable of analyzing guest review comments and predicting whether a guest is recommended or not. This assumption is based on the availability of true recommendations from hosts on selected stays, which serve as our training data. This task is akin to sentiment analysis. If you were tasked with developing such a classification system for Airbnb, the outcome would be a score assigned to each potential booking, representing the estimated probability of a guest being recommended or not, based on their previous reviews. The primary goal is to assist Airbnb hosts in effectively screening out undesirable guests, enabling hosts to decline bookings from guests with low scores.

However, an unexpected challenge arises when a racial discrimination lawsuit is filed against Airbnb. The plaintiffs allege that the sentiment analysis algorithm deliberately assigns low scores to specific ethnic groups, thus restricting their access to accommodations on the Airbnb platform.

As the lead engineer responsible for this algorithm, you may find yourself called as a witness in court. In this role, you would need to be well-prepared to explain your sentiment algorithm to both the judge and the jury. When presented with various reviews, you should be capable of elucidating why your algorithm assigns different scores to them.