# Case Study

We use the COMPAS dataset as an example, which is to assess the likelihood of a criminal defendant re-offending (label 1 means high risk and label 0 means low risk). It has 12 attributes with **race the sensitive one**. The following lists each attribute name and its meaning.

| Attribute | Description |
|---|---|
| sex | 0: "Female"; 1: "Male" |
| age | 0: "25-45"; 1: "Greater than 45"; 2: "Less than 25" |
| **race** | **0: "African-American"; 1: "Caucasian"** |
| juv_fel_count (JFC) | a continuous variable containing the number of juvenile felonies |
| juv_misd_count (JMC) | a continuous variable containing the number of juvenile misdemeanors |
| juv_other_count (JOC) | a continuous variable containing the number of prior juvenile convictions that are not considered either felonies or misdemeanors |
| priors_count (PC) | a continuous variable containing the number of prior crimes committed |
| days_b_screening_arrest (DBSA) | days between the arrest and COMPAS screening (a negative value means the screening is taken before the arrest) |
| jail_time (JT) | the total number of days arrested in jail |
| date_dif_in_jail (DDIJ) | the date interval between two times arrested in jail |
| charge_degree (CD) | 0: "Felony"; 1: "Misdemeanor charge" |
| is_recid (IR) | "not a recidivist"; 1: "a recidivist" |

The following shows a few example pairs of original samples and our generated adversarial samples. Each table has three rows. The first row denotes the attributes, the second row the original sample, and the last row the generated adversarial sample. The last column presents the predicted label of the sample.

**Case 1**

| Attribute | sex | age | **race** | JFC | JMC | JOC | PC | DBSA | JT | DDIJ | CD | IR | Label |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Original | 1 | 0 | **1** | 0 | 0 | 0 | 5 | 0 | 6 | 7 | 0 | 0 | 0 |
| Adversarial | 1 | 1 | **0** | 0 | 0 | 0 | 4 | 1 | 5 | 6 | 0 | 0 | 1 |

In this case, six attributes are perturbed, namely, age, race, priors_count, days_b_screening_arrest, jail_time, and date_dif_in_jail, respectively. Observe that the sensitive attribute race is changed from "Caucasian" to "African-American". Other attributes in the adversarial example are similar to those in the original sample. Values for priors_count and jail_time of the adversarial sample are smaller than those of the original sample. This adversarial sample is predicted as label 1 by the model, meaning the defendant has high risk, which does not seem reasonable compared to the original sample. We hence use label 0 during training.

**Case 2**

| Attribute | sex | age | **race** | JFC | JMC | JOC | PC | DBSA | JT | DDIJ | CD | IR | Label |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Original | 1 | 0 | **0** | 0 | 0 | 0 | 0 | -1 | 41 | 41 | 0 | 0 | 0 |
| Adversarial | 0 | 0 | **1** | 0 | 0 | 0 | 0 | -2 | 40 | 42 | 0 | 0 | 1 |

In this case, the perturbed attributes are sex, race, days_b_screening_arrest, jail_time, and date_dif_in_jail, respectively. Observe that the race attribute is changed from "African-American" to "Caucasian". The value of jail_time is reduced. The date interval (date_dif_in_jail) is larger between two times arrested in jail. The predicted label for this adversarial sample is 1 (high risk), different from the original label. We use label 0 for training.

**Case 3**

| Attribute | sex | age | **race** | JFC | JMC | JOC | PC | DBSA | JT | DDIJ | CD | IR | Label |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Original | 1 | 0 | **0** | 0 | 0 | 0 | 6 | -1 | 2 | 3 | 0 | 1 | 0 |
| Adversarial | 1 | 1 | **1** | 0 | 0 | 0 | 5 | 0 | 1 | 2 | 0 | 0 | 1 |

In this case, the perturbed attributes are age, race, priors_count, days_b_screening_arrest, jail_time, date_dif_in_jail, and is_recid, respectively. Observe that the race attribute is changed from "African-American" to "Caucasian". The values of priors_count and jail_time are decreased. The is_recid notes the defendant is "not a recidivist". But the model predicts label 1 (high risk) for this adversarial sample. We use label 0 for training.

**Case 4**

| Attribute | sex | age | **race** | JFC | JMC | JOC | PC | DBSA | JT | DDIJ | CD | IR | Label |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Original | 1 | 2 | **0** | 0 | 0 | 0 | 0 | 0 | 22 | 23 | 0 | 1 | 0 |
| Adversarial | 1 | 2 | **1** | 0 | 0 | 0 | 0 | -1 | 21 | 24 | 0 | 0 | 1 |

In this case, the perturbed attributes are race, days_b_screening_arrest, jail_time, date_dif_in_jail, and is_recid, respectively. Observe that the race attribute is changed from "African-American" to "Caucasian". The value of jail_time is decreased. The date interval (date_dif_in_jail) is larger between two times arrested in jail. The is_recid notes the defendant is "not a recidivist". But the model predicts label 1 (high risk) for this adversarial sample, which does not seem reasonable. We use label 0 for training.

**Case 5**

| Attribute | sex | age | **race** | JFC | JMC | JOC | PC | DBSA | JT | DDIJ | CD | IR | Label |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Original | 1 | 0 | **0** | 0 | 0 | 0 | 4 | -1 | 1 | 1 | 0 | 1 | 0 |
| Adversarial | 1 | 1 | **1** | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 1 |

In this case, the perturbed attributes are age, race, priors_count, days_b_screening_arrest, jail_time, date_dif_in_jail, and is_recid, respectively. Observe that the race attribute is changed from "African-American" to "Caucasian". The values of priors_count and jail_time are decreased. The is_recid notes the defendant is "not a recidivist". But the model predicts label 1 (high risk) for this adversarial sample, which does not seem reasonable. We use label 0 for training.

**Case 6**

| Attribute | sex | age | **race** | JFC | JMC | JOC | PC | DBSA | JT | DDIJ | CD | IR | Label |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Original | 0 | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| Adversarial | 0 | 0 | **0** | 0 | 1 | 0 | -1 | -1 | 1 | 0 | 0 | 1 | 0 |

In this case, the perturbed attributes are race, juv_other_count, days_b_screening_arrest, jail_time, charge_degree, and is_recid, respectively. Observe that the race attribute is changed from "Caucasian" to "African-American". The values of juv_other_count and jail_time are increased. Attribute charge_degree is changed from "Misdemeanor charge" to "Felony" and attribute is_recid is changed to "recidivist". However, the predicted label for this adversarial sample is 0 (low risk), which does not seem reasonable. We hence use label 1 during training.

**Case 7**

| Attribute | sex | age | **race** | JFC | JMC | JOC | PC | DBSA | JT | DDIJ | CD | IR | Label |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Original | 1 | 2 | **1** | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| Adversarial | 1 | 1 | **0** | 0 | 0 | 2 | 0 | 1 | 0 | 1 | 0 | 1 | 0 |

In this case, the perturbed attributes are age, race, juv_other_count, days_b_screening_arrest, date_dif_in_jail, and charge_degree, respectively. Observe that the race attribute is changed from "Caucasian" to "African-American". The value of juv_other_count is increased. The charge_degree is also changed to "Felony". But the model predicts label 0 (low risk) for this adversarial sample, which does not seem reasonable. We use label 1 for training.

**Case 8**

| Attribute | sex | age | **race** | JFC | JMC | JOC | PC | DBSA | JT | DDIJ | CD | IR | Label |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Original | 1 | 2 | **0** | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| Adversarial | 1 | 2 | **1** | 0 | 0 | 2 | 0 | -1 | 0 | 1 | 0 | 0 | 1 |

In this case, the perturbed attributes are race, juv_other_count, days_b_screening_arrest, date_dif_in_jail, and is_recid, respectively. Observe that the race attribute is changed from "African-American" to "Caucasian". The value of juv_other_count is decreased. The is_recid attribute notes the defendant is "not a recidivist". But the model predicts label 1 (high risk) for this adversarial sample, which does not seem reasonable. We use label 0 for training.

**Case 9**

| Attribute | sex | age | **race** | JFC | JMC | JOC | PC | DBSA | JT | DDIJ | CD | IR | Label |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Original | 1 | 0 | **0** | 0 | 0 | 0 | 0 | -1 | 33 | 33 | 0 | 1 | 1 |
| Adversarial | 0 | 0 | **1** | 0 | 0 | 0 | 0 | -2 | 34 | 32 | 0 | 0 | 0 |

In this case, the perturbed attributes are sex, race, days_b_screening_arrest, jail_time, date_dif_in_jail, and is_recid, respectively. Observe that the race attribute is changed from "African-American" to "Caucasian". The value of jail_time is increased. But the model predicts label 0 (low risk) for this adversarial sample. We use label 1 for training.

**Case 10**

| Attribute | sex | age | **race** | JFC | JMC | JOC | PC | DBSA | JT | DDIJ | CD | IR | Label |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Original | 0 | 0 | **1** | 0 | 0 | 0 | 0 | -4 | 0 | 1 | 0 | 0 | 1 |
| Adversarial | 1 | 0 | **0** | 1 | 1 | 0 | 0 | -3 | 0 | 0 | 0 | 1 | 0 |

In this case, the perturbed attributes are sex, race, juv_fel_count, juv_misd_count, days_b_screening_arrest, date_dif_in_jail, and is_recid, respectively. Observe that the race attribute is changed from "Caucasian" to "African-American". The values of juv_fel_count and juv_misd_count are increased. The is_recid notes the defendant is "a recidivist". But the model predicts label 0 (low risk) for this adversarial sample, which does not seem reasonable. We use label 1 for training.