FAIRNESS AND BIAS IN PREDICTIVE AI MODELS

Dan IANCU

daniancu@stanford.edu

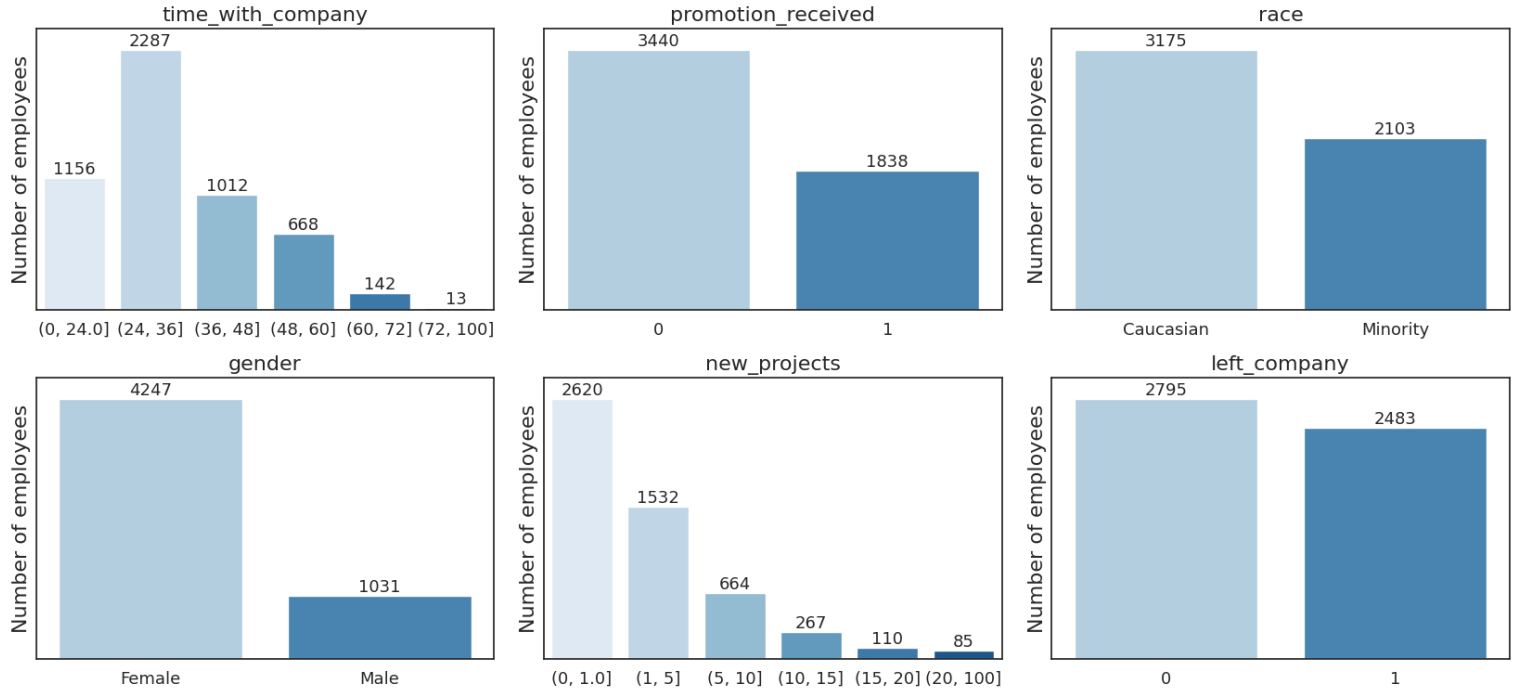
Hands-on Exercise

- Dataset on employee turnover
  - 5,278 company employees from different locations
  - We have information on each employee recorded at some point in the past, and we also know whether they left the company during the subsequent year
  - This is a **real dataset**, but from a very different context (to be revealed in class!)

- **Goals:**
  1. **Understand the data** and identify potential **sources of bias**
  2. Build our own interpretable AI model – a **Decision Tree** – and assess it for bias
  3. Evaluate **a proprietary/black-box model** for bias

Part 1: Understand the data

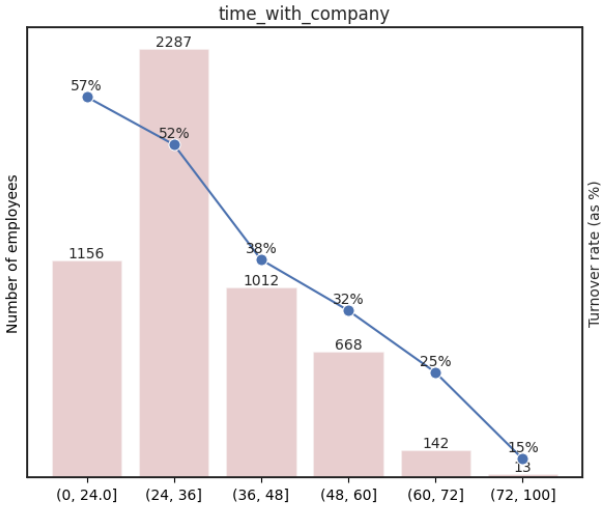
- There are 6 data features available
  - ***time_with_company*** *:* the employee's tenure with the company (in months)
  - ***promotion_received*** *:* whether the employee was promoted during prior 2 years (**1** or **0**)
  - ***race*** *:* two values in data, **Caucasian** or **Minority**
  - ***gender*** *:* two values in data, **Male** or **Female**
  - ***new_projects*** *:* total number of new projects the employee was involved in during prior 2 years
  - ***left_company*** *:* whether the employee left the company in the subsequent year (**1** or **0**)

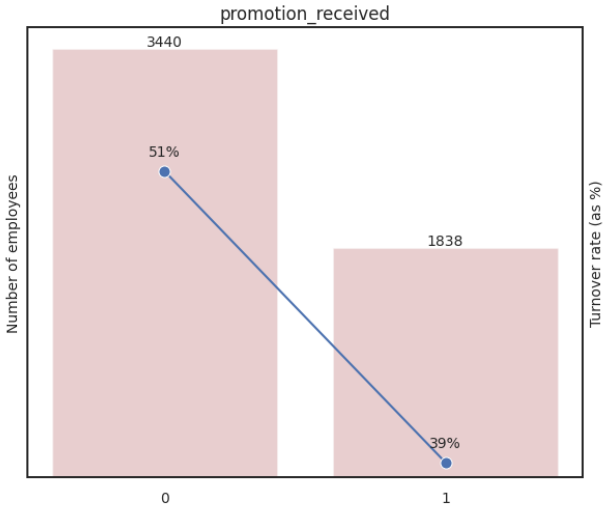


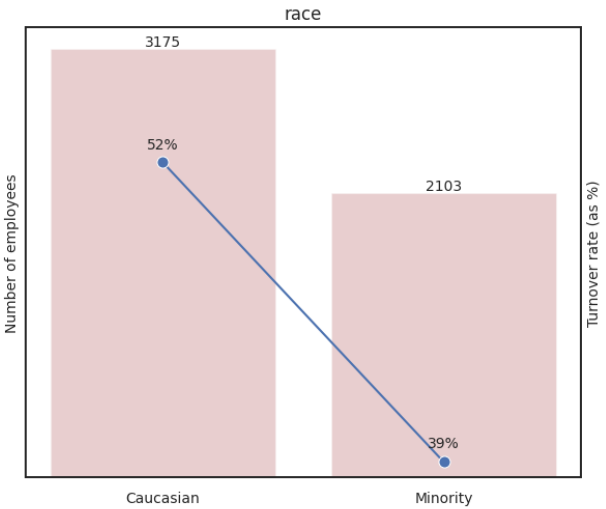
Part 1: Relationship with turnover

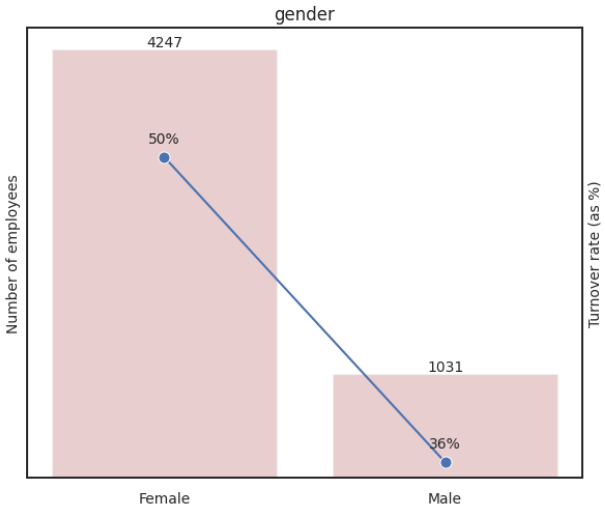
Q1: How would you characterize the relationship between turnover and each data feature?

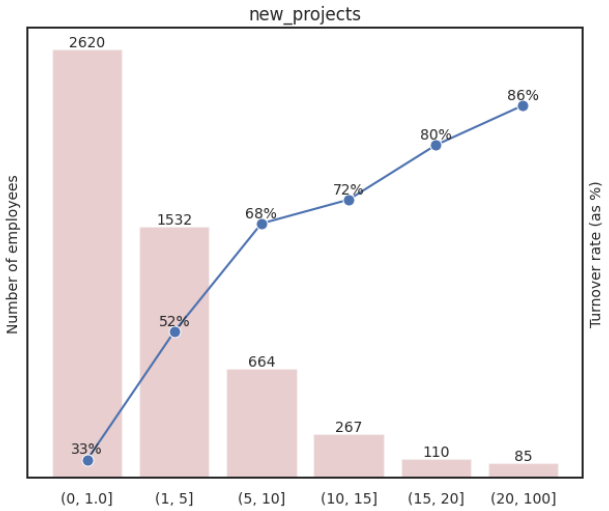
(please answer in the poll)











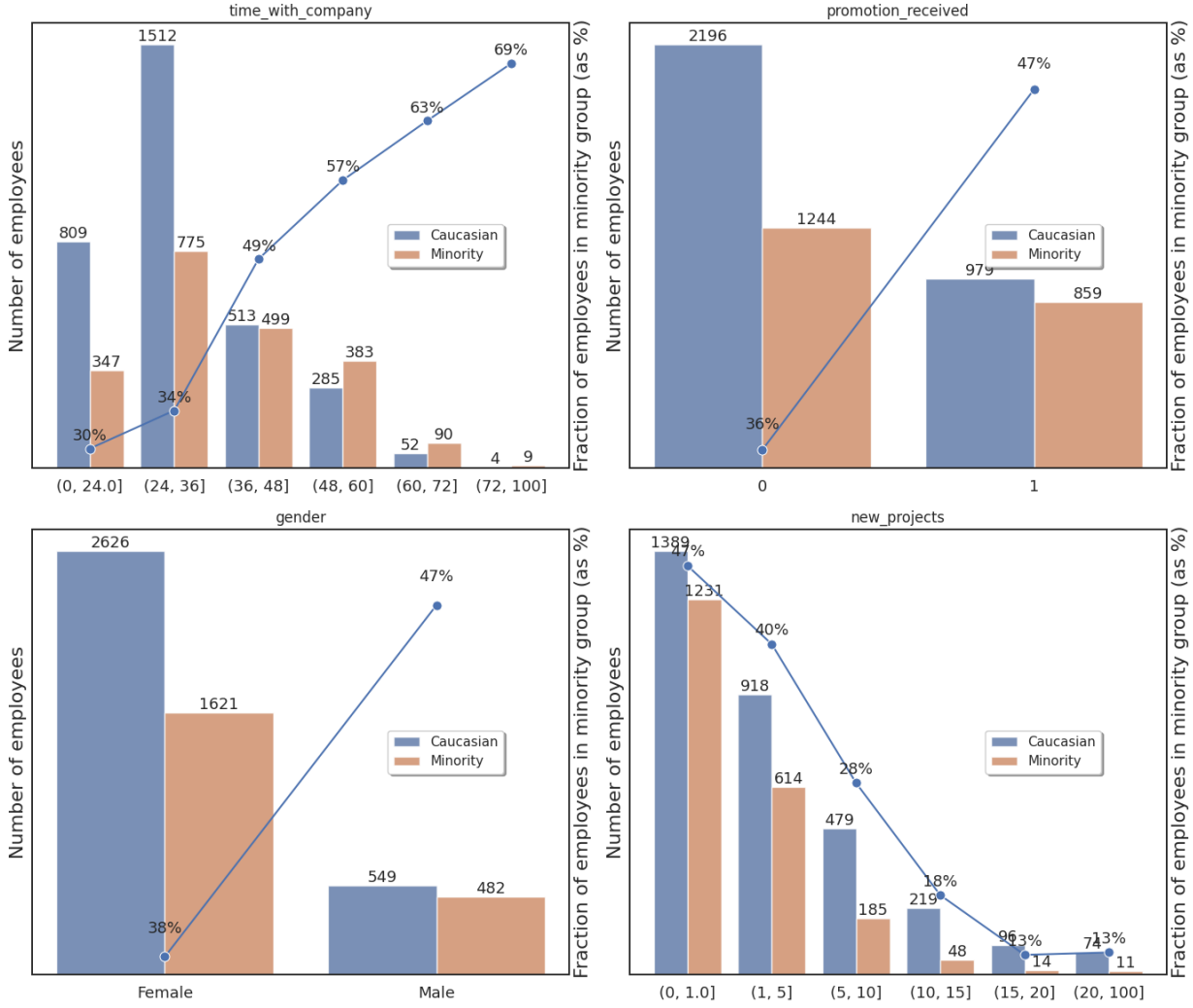
Clarification about each plot:

- The **pink bars** show the **number of employees** in each category (that value is on top of each bar).
- The **blue line plot** shows the **turnover rate,** calculated as the fraction of employees in that category who left the company within a year.

For example: *in the plot corresponding to* time_with_company*, among the 1,156 employees with tenure of at most 24 months, the turnover rate was 57%.*

Part 1: Relationship with race

How would you characterize the relationship between race and the other data features?
(brainstorm)



Clarification about each plot.

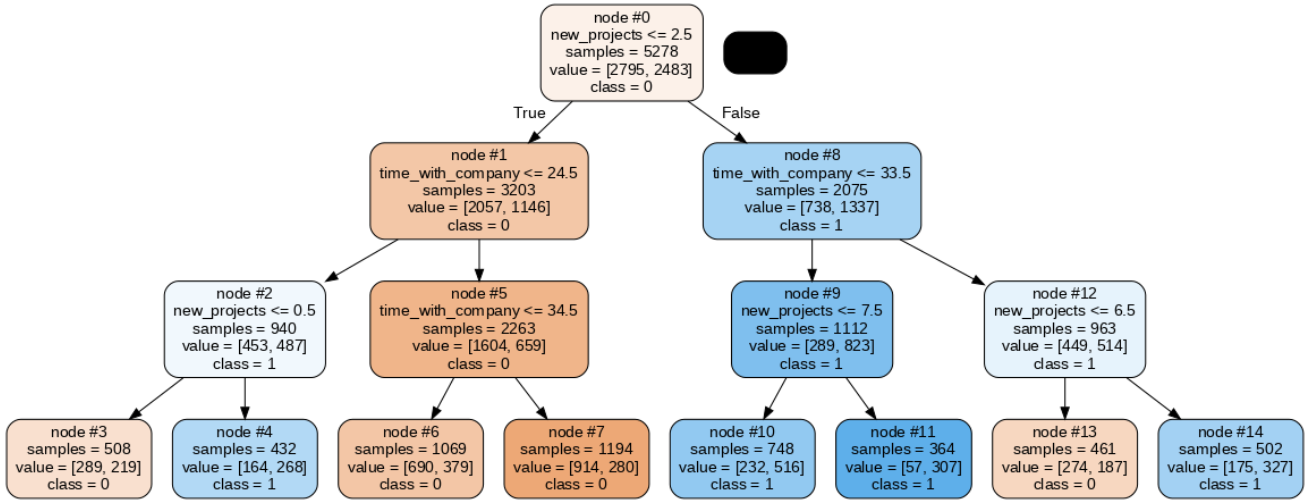
- The **bars** show the **number of employees** in each category (the value on top of each bar).
- The **blue line plot** shows the **fraction of employees** within each category **who are a minority** group**.**

Example: *in the plot for* ***time_with_company****, among the employees with tenure of 0-24 months, 809 are Caucasian and 347 (i.e., 57%) are a Minority group.*

Part 2: An Interpretable AI Model

Q2: Is this model exhibiting racial bias? *(please answer in the poll!)*

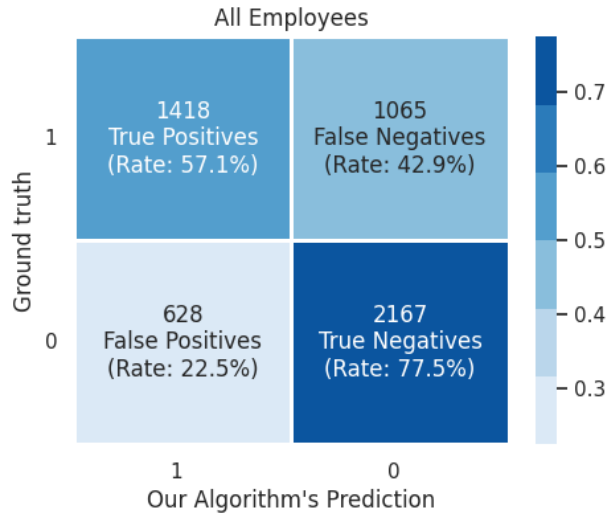
Here, we train our own AI model with all the data, including features like **race** and **gender.**

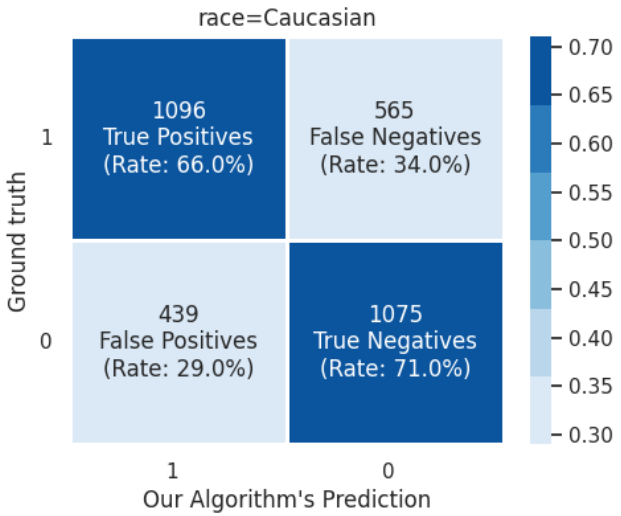


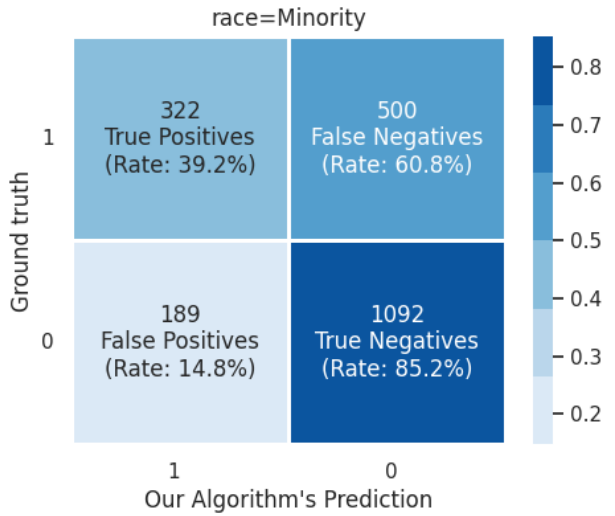
Clarification about visualizations.

- **Left:** you can see the way the model works, i.e., what features it uses and how it makes predictions.
- **Below:** you can see the model's performance on the data, calculated with the confusion matrix.

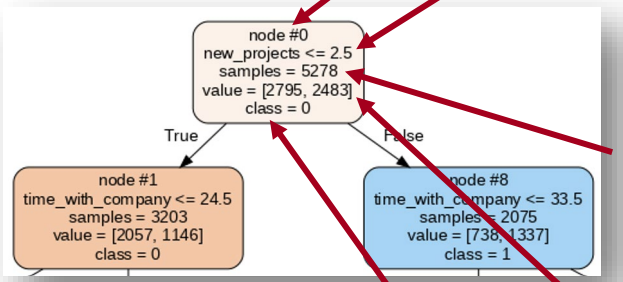
For help interpreting these outputs, see the next two pages!







How To Interpret a **Decision Tree**?



- "node #0" is a **unique identifier** for the node
- second line has a logical condition comparing a data **feature** with a certain **threshold**; the left subtree contains all the data where the condition is "**True**" and the right subtree contains the rest ("**False**")

  *Example: in node #0, we check if "*new_projects <= 2.5*". All data points that satisfy this are the placed in node #1, and the rest of the data are in node #8.*
- **samples** indicates how many data points fall in that node

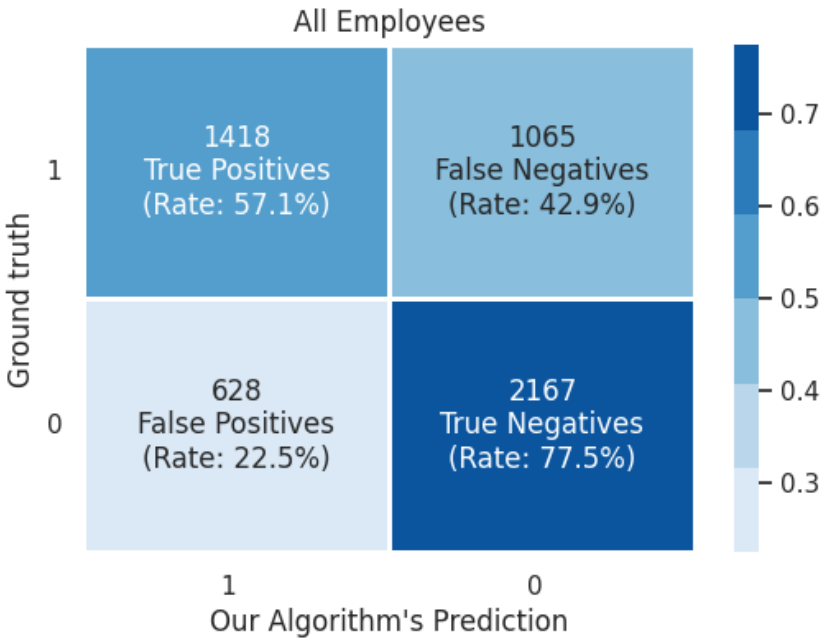
  *Example: node #0 contains 5,278 samples (the entire data), whereas node #1 contains 3,203 samples (recall that node #1 as all the samples satisfying "*new_projects <= 2.5*")*
- **value** indicates how many data points take value 0 and 1 for the predicted target, respectively

  Example: in node #0 we have 2,795 samples with left_company=0 *and 2,483 samples with* left_company=1. *Note that the sum of these two numbers always equals the number reported under "***samples***" on the line above (for node #0, that would be 5,278).*
- **class** indicates which value occurs most often for the predicted target. This is also indicated by the color-coding of nodes: orange means majority 0 and blue means majority 1, and the deeper the color the heavier the majority.

  E.g., in node #0 we have more data with left_company=0 *(namely, 2,438 samples) than with* left_company=1 *(namely, 2,483), so class=0 indicates that the majority is 0, and the node is colored in a light shade of orange, which indicates this is not a heavy majority*

How To Interpret a **Confusion Matrix**?

- The quality of predictions is summarized with a **confusion matrix**:



Ground truth: what happened in reality
Prediction: what the algorithm predicted/thought would happen

True: prediction matches reality (the algorithm is right)
False: prediction does not match reality (the algorithm is wrong)
Positive: prediction = 1
Negative: prediction = 0

The **rates** are calculated as a fraction of the total on the row.

***Example**. There are 1,418 true positives, meaning this is the number of cases when the algorithm predicts an employee would leave the company (Prediction=1) and that employee actually leaves the company (Ground Truth=1). Because the total number of employees who actually leave the company is 1,418 + 1,065, the rate of true positives is 1,418 / (1,418+1,065) = 57.1%.*

Part 3: Evaluating A Black-Box Model – RETAIN

Suppose that your company is using a risk scoring tool called **RETAIN**, which is a proprietary software designed by a third-party provider. The tool can produce, for every employee, **a risk score from 1 to 10** (with higher values indicating higher risk) and a **risk label** based on the score: those with risk score of 5 or above are labeled as **1** (high risk), and those with scores of 4 or below are labeled **0** (low risk).

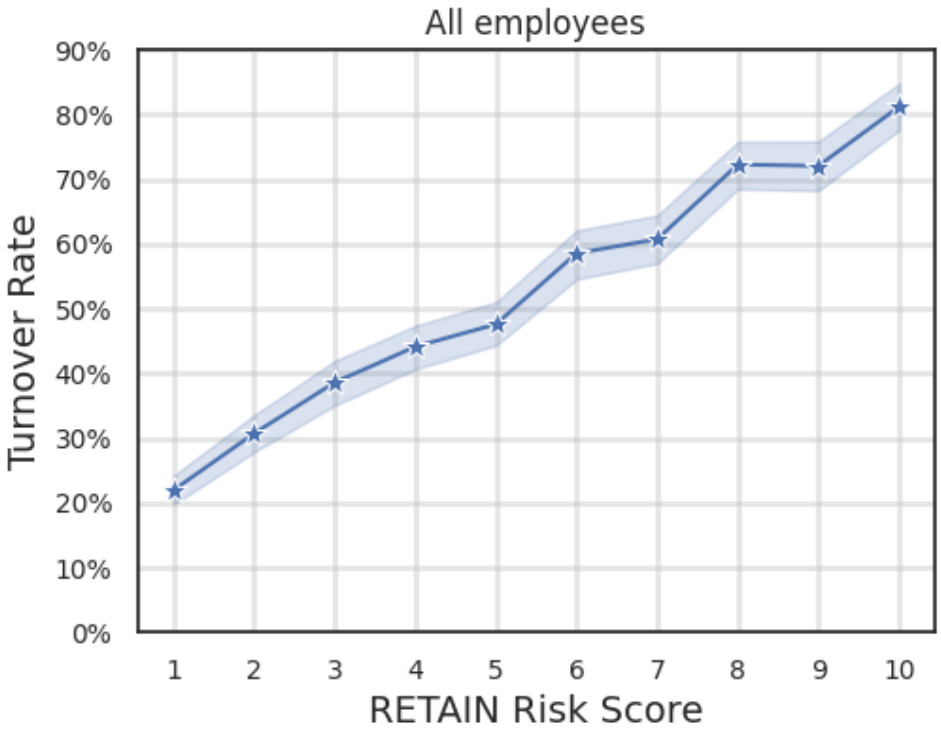
The software provider assures you that **RETAIN was created without any protected features like gender or race.** You cannot understand how RETAIN reaches its predictions, but you can use it to construct risk scores for all the employees in your data and evaluate the relationship between RETAIN's predictions and the ground truth (i.e., what happened with your employees in reality). The next pages do exactly that. Using these outputs, please address the following questions:

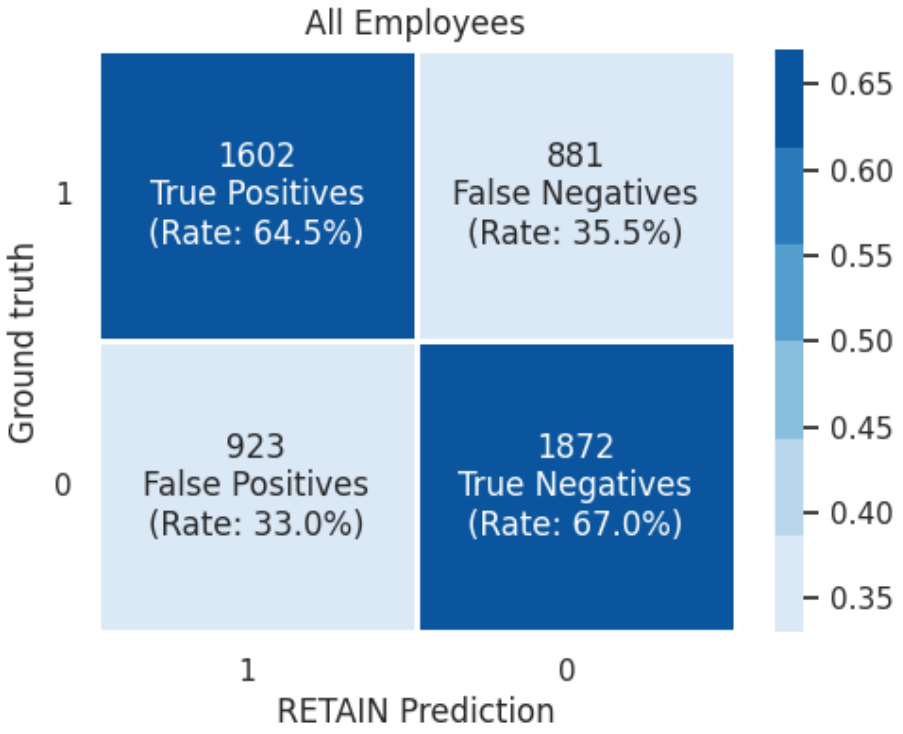
Q3. Suppose you used the RETAIN algorithm to make decisions and you completely ignored the gender feature. Would your decisions be gender biased? If so, against which gender? ***(please answer in the poll)***

Q4. Suppose you used the RETAIN algorithm to make decisions and you completely ignored the race feature. Would your decisions be racially biased? If so, against which race? ***(please answer in the poll)***

Q5. If there is a bias, how could you fix it? ***(brainstorming)***

RETAIN predictions in the entire dataset

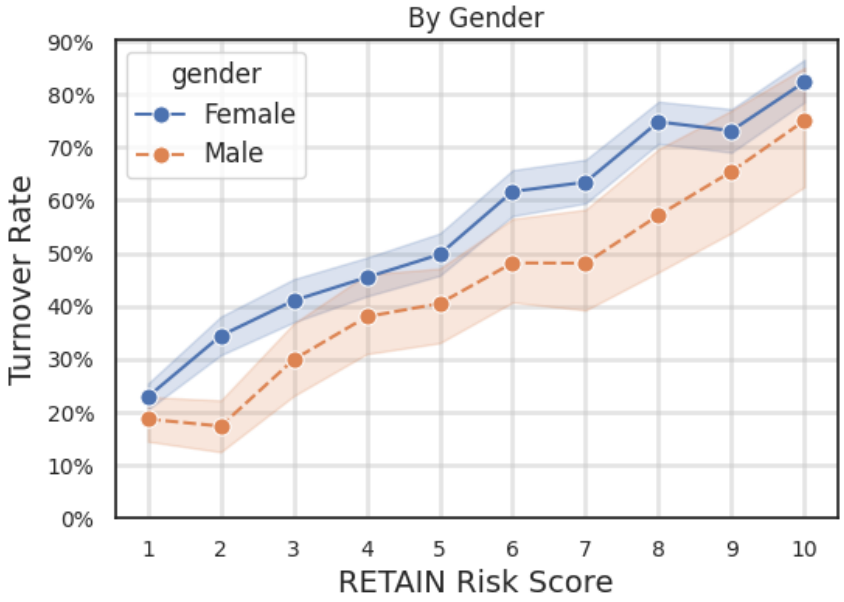




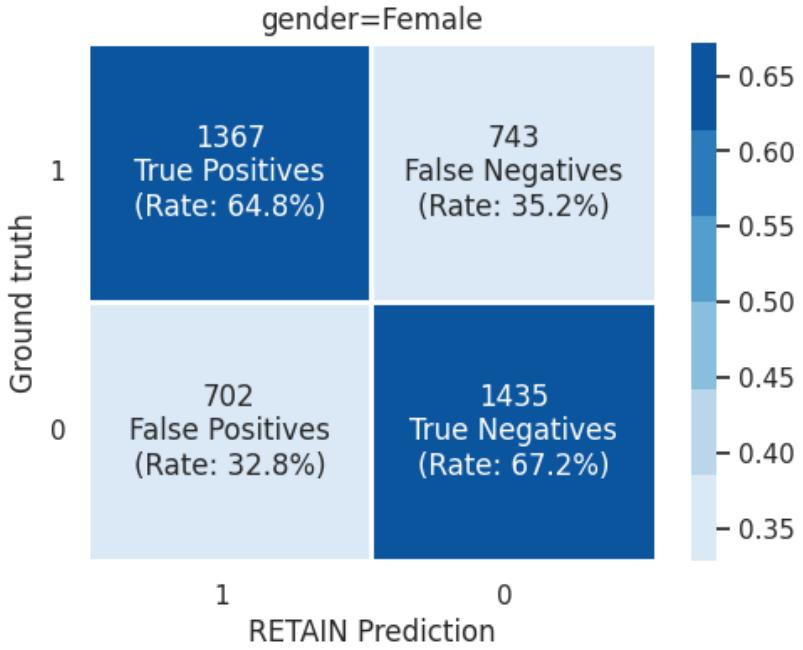
A clarification about the figures.

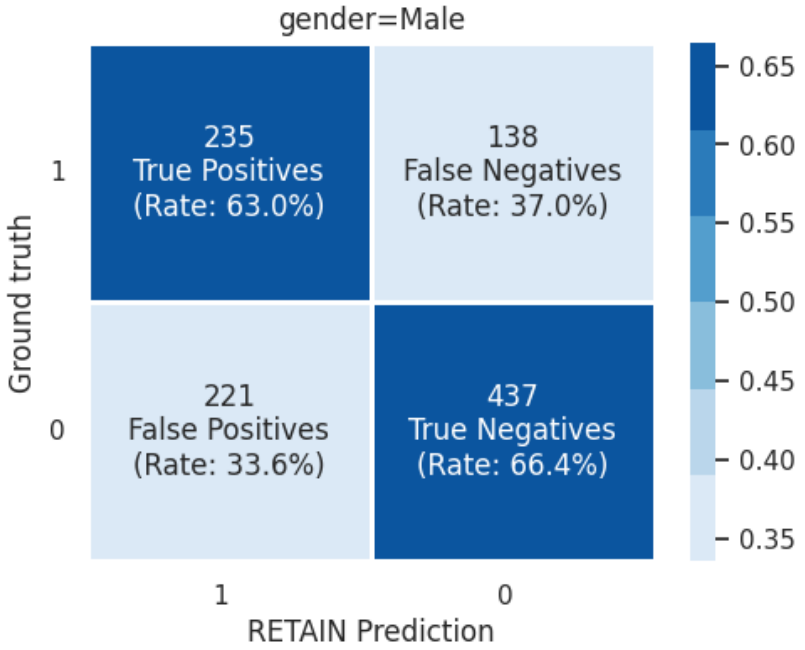
• The **left figure** shows the **true turnover rate** as a function of the **RETAIN risk score**. The center line indicates the turnover rate for all employees with the given score, and there is also a confidence band.

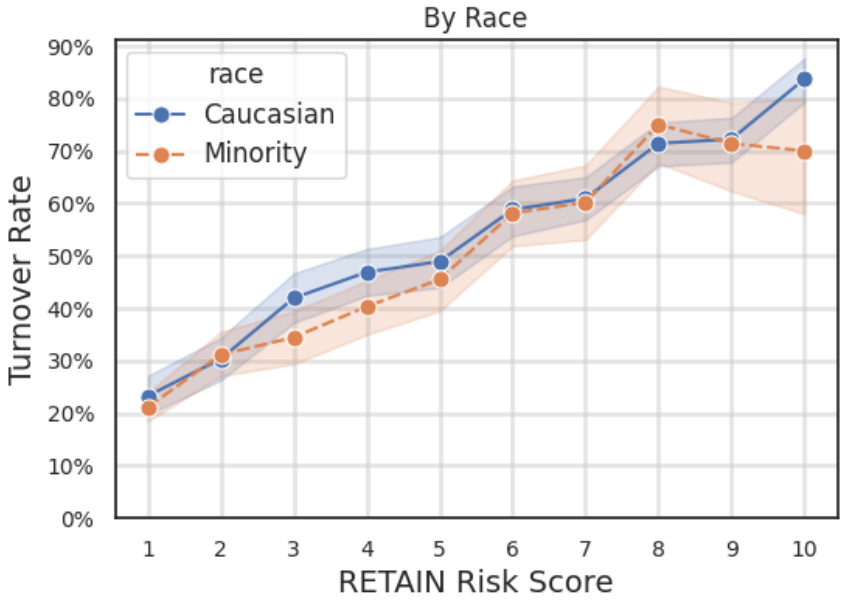
• The **right figure** shows a **confusion matrix** obtained based on the **RETAIN risk label**.



RETAIN predictions by **gender**







RETAIN predictions by **race**