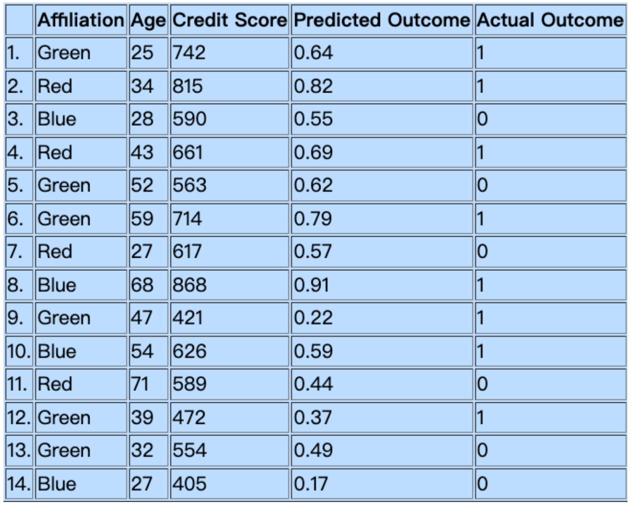
Quiz 10 Solutions

CSE 4/574

Fall, 2024

Question 1

You have been given the training data and real-valued outcomes of a model trained to predict loan repayment. An outcome of 1 means that the person successfully paid back their loan, where an outcome of 0 means that the person defaulted. The affiliation feature is a protected attribute. Given these data and results, which of the following is true?



	Affiliation	Age	Credit Score	Predicted Outcome	Actual Outcome
1.	Green	25	742	0.64	1
2.	Red	34	815	0.82	1
3.	Blue	28	590	0.55	0
4.	Red	43	661	0.69	1
5.	Green	52	563	0.62	0
6.	Green	59	714	0.79	1
7.	Red	27	617	0.57	0
8.	Blue	68	868	0.91	1
9.	Green	47	421	0.22	1
10.	Blue	54	626	0.59	1
11.	Red	71	589	0.44	0
12.	Green	39	472	0.37	1
13.	Green	32	554	0.49	0
14.	Blue	27	405	0.17	0

Correct Choice

The thresholds red=0.67, green=0.51, blue=0.58 satisfy demographic parity.

Explanation:
Using this threshold setting, the probability of predicting positive for red, green and blue affiliations are all 0.5, therefore satisfying the demographic parity.

Question 2

Which of these is not an example of adversarial attacks on an AI system?

Correct Choice

Pretend to be the attacker, generate a number of adversarial examples against your own network, and then explicitly train the model to not be fooled by them.
Explaination:
In the scenario, the intent is called adversarial training, which is constructive. The adversarial examples are generated by the model's developers to strengthen the model's robustness, not to deceive or undermine its functionality. Additionally, this action is not outside the development or deployment environment. However, the rest of the choices are all malicious actors that deliberately attempt to subvert the functionality of AI systems.