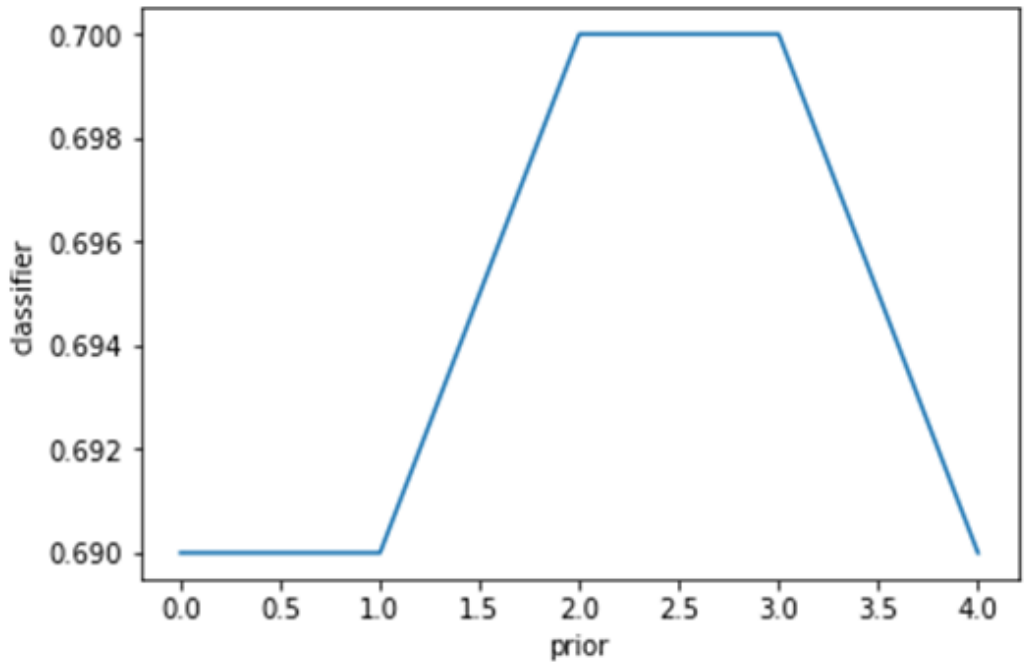
Project 3

CSE 474
Group:23

Anthony Rubin
Vikram Singh
Abhijeet Verma

1.) The impact of priors on the performance of the classifier



Explanation: We can clearly see from the graph that with change in the value of prior, value of the classifier changes a little. Which conclude that probability belongs to the good credit or bad credit depends on the prior value.

2.)Disparate Impact score of the cross-validation predictions of the Naive Bayes classifier

DI =P(Y = 2|S != 1)/P(Y = 2|S = 1)

sensitive features - gender

Attribute 1: Gender

1 : male

2 : female

si=0 and p=0.5

Disparate Impact Score for Sensitive Feature 0 = 1.10

Explanation: As we can see that the disparate impact score is 1.10 for the y=2 and s is not equal to 1 i.e “females”.The values is higher for the numerator which proves that the

impact is greater to females and therefore is more biased towards the males.The sensitive feature has more “bad” credit to the females than to males.Therefore the classifier assigns more bad credit to females or the “unprivileged class”

sensitive features - age

Attribute 2: Age

1 : older than 25

2 : younger than 25

si=1 and p=0.5

Disparate Impact Score for Sensitive Feature 1 = 1.25

Explanation: As we can see that the disparate impact score is 1.25 for the y=2 and s is not equal to 1 i.e “younger than 25”.The values is higher for the numerator which proves that the impact is greater to people younger than 25 and therefore is more biased towards the people older than 25.The sensitive feature assigned more “bad” credit to the people younger than it assigned to older people.Therefore the classifier proves to be assigning more bad credit for the people younger than 25 than to people older than 25.

3.) Relationship between p and fairness of the classifier.

For gender

si=0 and p=0.2

Disparate Impact Score for Sensitive Feature 0 = 0.41

si=0 and p=0.4

Disparate Impact Score for Sensitive Feature 0 = 0.73

si=0 and p=0.5

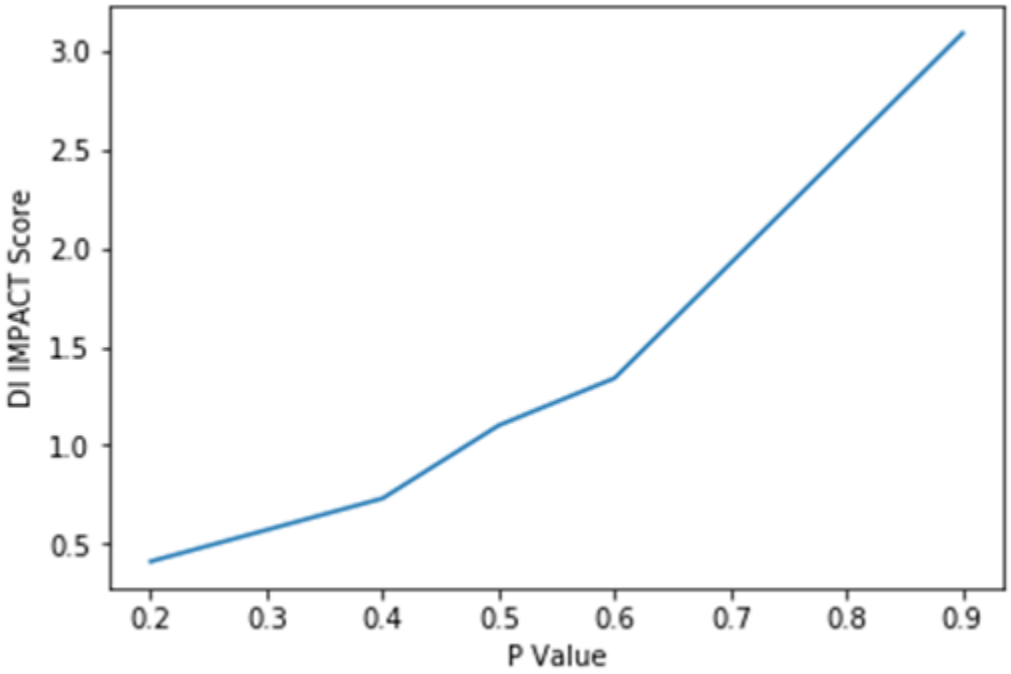
Disparate Impact Score for Sensitive Feature 0 = 1.10

si=0 and p=0.6

Disparate Impact Score for Sensitive Feature 0 = 1.34

si=0 and p=0.9

Disparate Impact Score for Sensitive Feature 0 = 3.09



Explanation:

The values grows for the disparate impact as the values of p is increased.The graph can be explained for p>0.5,p<0.5 and p=0.5.

For p=0.5 we can see that there exist no artificial bias towards the score and revolves around one.That means the sensitive feature does not make any impact towards it.That means the ratio of bad credits being assigned to the males and females is same.

For p<0.5 the data is biased against the privileged people as the denominator increases as p decreases making it more biased as assigning of the bad credit to the privileged people or for s==2.Therefore we are inducing artificial bias to work for the unprivileged people.The artificial bias is working for the females and assigning more bad credits to the males.

For p>0.5 the data is biased towards the privileged people as the numerator increases as p increases making it more biased as assigning of the bad credit to the unprivileged people or for s==1.Therefore the classifier is assigning more bad credits to the unprivileged class i.e females and acting against it.

For age

si=1 and p=0.2
Disparate Impact Score for Sensitive Feature 1 = 0.31

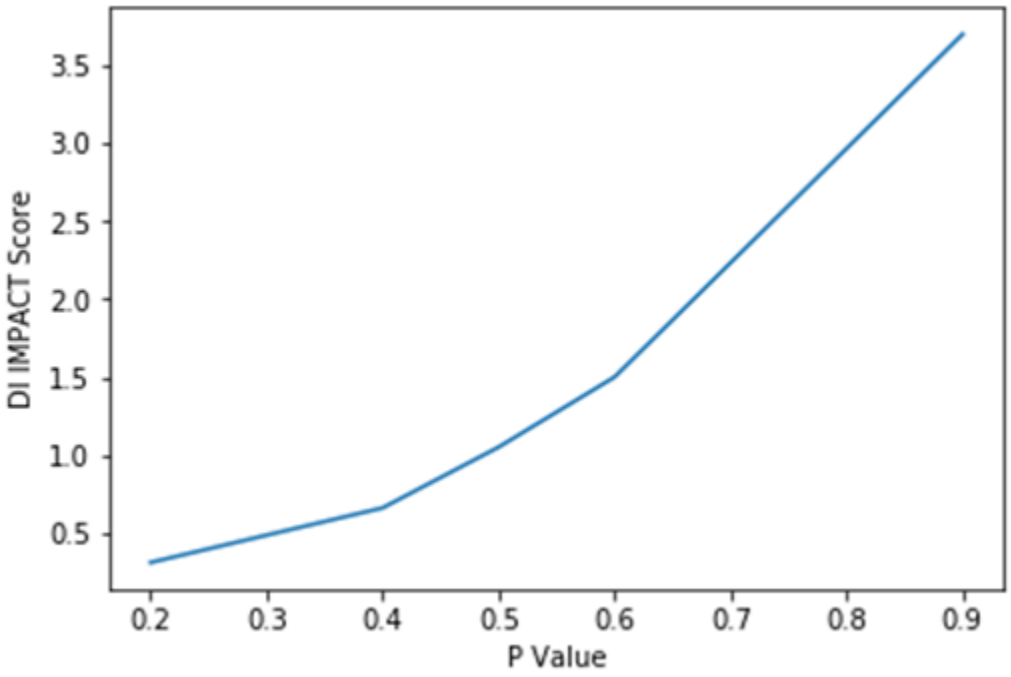
si=1 and p=0.4

Disparate Impact Score for Sensitive Feature 1 = 0.66

si=1 and p=0.5
Disparate Impact Score for Sensitive Feature 1 = 1.05

si=1 and p=0.6
Disparate Impact Score for Sensitive Feature 1 = 1.50

si=1 and p=0.9
Disparate Impact Score for Sensitive Feature 1 = 3.70



Explanation:

The values grows for the disparate impact as the values of p is increased.The graph can be explained for p>0.5,p<0.5 and p=0.5.

For p=0.5 we can see that there exist no artificial bias towards the score and revolves around one.That means the sensitive feature does not make any impact towards it.That means the ratio of bad credits being assigned to the people younger and older than 25 is same.

For p<0.5 the data is biased against the privileged people as the denominator increases as p decreases making it more biased as assigning of the bad credit to the privileged people or for s==2.Therefore we are inducing artificial bias to work for the unprivileged people.The artificial bias is working for the “people younger than 25” and assigning more bad credits to the “older people than 25”.

For p>0.5 the data is biased towards the privileged people as the numerator increases as p increases making it more biased as assigning of the bad credit to the unprivileged people or for s==1.Therefore the classifier is assigning more bad credits to the unprivileged class i.e “people younger than 25” and acting against it.Therefore making it more biased against the unprivileged customers.