Is AI Necessary In Deciding Whether an Offender Is Likely To Recidivate, With & Without the Effect of Protected Characteristics Final Paper

Abstract:
Our research topic addresses the issue of whether or not an AI algorithm can be developed to forecast whether or not a criminal would recidivate by removing protected characteristics from models in terms of both accuracy, precision and recall and their equality across different groups, as well as if such a tool is necessary. The background of this issue is that there is always an underlying human prejudice present in the criminal justice system when it comes to making judgments, thus it is crucial to determine whether or not this human bias is required while making these decisions and whether there is similar kind of bias with machine learning methods. There are various ways to approach and respond to this topic, but the strategy we used involved examining the key variables that might influence prejudice, such as gender or ethnicity, and comparing how well the AI system can predict whether someone is likely to recidivate with or without incorporating these protected characteristics. The research's significant findings are that, after utilizing a number of different models, logistic regression was used to determine whether or not an offender was likely to recidivate. The accuracy of the predictions for our logistic regression model was 67.74 percent without the removal of any protected characteristics that hold risk of being discriminated against unfairly, but it increased to 68.16 percent when race and gender were taken into account. The random Forest model revealed that, after accounting for the gender characteristic, the mean absolute error—the percentage of variance between each measurement and the quantity's actual value—was 34.06 percent along with an accuracy score of 65.94 percent. The main results drawn from this demonstrate that utilizing various models and measuring the precision, recall, and accuracy of each race and gender are comparable with or without protected characteristics. Although the accuracy of this still isn't the finest, it could demonstrate that bias may not be as significant a factor in decision-making as it may first appear. This, however, can also be contested because the dataset utilized to train these models could be encoding real world bias such that race and gender get encoded into other variables such as prior offenses; this would mean that even if we removed race and gender data it would still be present in our model through other variables. The question of whether or not we should remove protected characteristics is required and needed remains unanswered as a result.

Introduction:

Our research question looks at how AI can be used to determine how likely it is for an offender to commit recidivism, as well as looking at how to increase the accuracy of these results without the affect and impact of protected characteristics from different areas such as gender and race. The underlying problem at hand is that when looking at how likely someone is to reoffend along with appropriate sentencing, there lies a lot of subconscious bias that affects decisions made, which we want to try to minimize. The importance of this research problem is that it presents the ability to have an AI algorithm that can accurately predict and determine how likely it is for an offender to commit recidivism and being able to do that without the bias in different aspects, as bias always plays a role, whether or not it is prominent, in decisions made within the criminal justice system and that has had an impact throughout the years and it is important to find ways to minimize this bias, as well as ensuring the accuracy The methods that we had carried out was training and testing out the data is taking out the columns and factors that can play a role in bias, mainly gender and race, and trying different models to try and see which one can provide the best accuracy when looking at how likely someone is to commit recidivism. The results reached were unexpected and resulted in a different outcome to what was hoped. The models used are logistic regression, neural networks and Random Forest Classifier.

Background:

Since lately, artificial intelligence (AI) has been employed extensively around the world, with the COMPAS algorithm and Core Criminal Justice Research Literature Databases being a few examples. Using AI in the criminal justice system enables the criminal justice system to keep track of all offenders and uses this information to decide a variety of different goals. Some may state that the application of AI in this subject is crucial because it enables a more thorough investigation into the question of whether or not prejudices within the criminal justice system can be confirmed, yet others remain cautious that AI gives the appearance of fairness while making it very hard to have accountability or transparency with regard to how legal decisions are being made..

Current Approaches to Reducing Limitations for how Likely an Offender is to Recidivate:

This research question was approached in one way using a more technical approach. A machine learning strategy is suggested in the study titled "Classification of Criminal Recidivism Using Machine Learning Techniques" to identify and forecast a criminal's propensity for recidivism. The suggested approach aids in categorizing criminals into groups with Low, Medium, and High recidivism risks. By overcoming the restrictions of current law enforcement practices, a machine learning model seeks to address one of the aforementioned causes. This will allow the authorities to decide whether to grant parole to criminals who may commit repeat offenses in a reasoned, statistically and analytically sound manner. When these techniques are employed in parole hearings, Mehta et al claims to classifying criminals with accuracy of 87.81% using Random Forest, 86.88% using the K-nearest Neighbors algorithm, and 75.29% using Logistic Regression helps lower crime (Mehta, Shah, Patel and Kanani, 2020). According to the accuracy score, a Random Forest Classifier would produce the best results for each offender with a given set of attributes, although this differs from the results made throughout my experiments in which the Random Forest Classifier gave similar results to other models, yet was

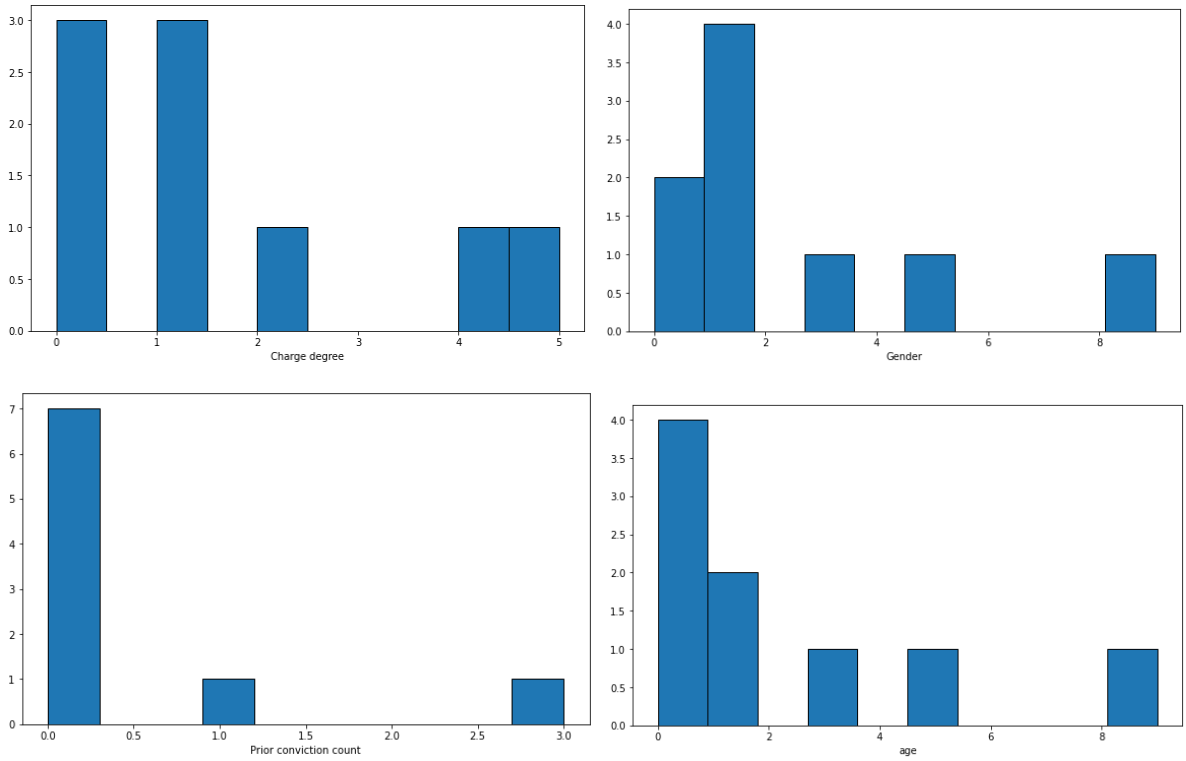
slightly worse compared to the other two models. The approach should, however, be generalized for other regional prisons/facilities and should be undertaken over a longer length of time to acquire higher accuracy, which is one of the drawbacks of this. In the coming years, it should also aid in lowering criminal recidivism. Additionally, real-world biases are stored in the data, but thanks to "cold, hard facts," this perception may now be hidden (Mehta, Shah, Patel and Kanani, 2020).

The article "How we Analyzed the COMPAS recidivism algorithm" offers yet another non-technical method. At nearly the same rate, the system accurately predicted recidivism for black and white defendants, but it made errors in very different ways. White defendants were more likely to be misclassified as low risk than black defendants, and vice versa was true for the reverse (Angwin, Larson, Kirchner and Mattu, 2016). The risk of violent recidivism is predicted using the COMPAS program. Black defendants are more likely than white defendants to obtain a better score, and young defendants are more likely to do so than middle-aged defendants. White offenders recidivated at a greater rate than black defendants, according to the COMPAS recidivism score's predictive accuracy. In Broward County, Florida, they examined more than 10,000 criminal defendants and discovered that the COMPAS tool properly predicted recidivism 61% of the time, but just 20% of the time for violent recidivism (Angwin, Larson, Kirchner and Mattu, 2016). Researchers looked at 19 various recidivism risk approaches utilized in the US and found that their predictive validity was, at best, modest (Angwin, Larson, Kirchner and Mattu, 2016). They were unable to locate a sizable collection of American research that looked at whether risk assessments were racially biased. 42.7 percent of African Americans were mistakenly categorized as high risk, compared to 27.7 percent of Caucasians and 25 percent of Hispanics, according to a study of 532 male residents of a work-release program (Angwin, Larson, Kirchner and Mattu, 2016). To test the idea that race is a factor in COMPAS scores, they created a logistic regression model. They used factors such as age, race, and gender to model the odds of getting a higher COMPAS score. We found that defendants younger than 25 were 2.5 times as likely to get a higher score than middle aged offenders (Angwin, Larson, Kirchner and Mattu, 2016). To conclude, They found no significant difference in the hazards of high and low risk black defendants and high and low risk white defendants using COMPAS's violent recidivism score. A black defendant is more likely to be incorrectly classified by the algorithm as being at a higher risk than a white defendant, while a white defendant is more likely to be predicted incorrectly as not committing more crimes upon release. Regardless of the COMPAS score, it was discovered that black defendants were more likely than white defendants to get bogus high risk ratings. White defendants were misclassified as having a low risk of violent recidivism 63.2 percent more frequently than black defendants, while black defendants were misclassified as having a greater risk of violent recidivism twice as frequently as white defendants (Angwin, Larson, Kirchner and Mattu, 2016).

Dataset:
The dataset we are using contains a list of offenders that explains how many felony and misdemeanor charges they have and how many prior convictions they have, as well as the current charge they have and the description of their charge. It also contains some information about the offenders age, gender and race. This allows more information to use for the algorithm which can allow a more accurate training of the dataset in multiple aspects of reducing bias. This dataset originally had 42 one-hot columns, that were boolean variables for 9 different categories. We were able to decrease the number of columns to 9 after recoding it to combine some of the columns since doing so would make it easier to work with and more effective because some of them were for the same category. This dataset consists of two sets, each with the same number of columns both before and after recoding, but with different numbers of samples. We employed a 70 to 30 training to testing

ratio split to our model. There were a total of 9 distinct categories created by combining the categories that were similar. These categories—juvenile felony count, juvenile misdemeanor count, and juvenile other count—define the types of offenses the offender has committed; misdemeanors are less serious than felonies and other offenses are distinct from them. The prior conviction count column indicates whether the offender has ever been convicted before the current offense. The charge degree and charge description columns follow, and these let us know what sort of offense they are now facing as well as a very brief explanation of it, such as "drug connected" or "no charge" if they are not facing any charges. The age, gender, and race of the criminal are revealed in the last three columns. Using the training data we were provided, we pre-processed the data to gain a sense of what kind of data we were working with. Our data was already divided into separate sets in the dataset we utilized, thus we didn't need to do that (See diagram below).Then, in order to display the age, gender, race, and number of prior convictions, we created a histogram.

```
gender mean:  2.3333333333333335
gender standard deviation:  2.7888667551135846
race mean:  2.0
race standard deviation:  3.5276684147527875
charges mean:  1.5555555555555556
charges standard deviation:  1.7069212773041353
```



In order to attempt to visualize the association between gender and the number of past convictions, we also developed a scatterplot diagram, with the color indicating the race of each participant. The fact that there were so few plots suggested that many of them were stacked on top of one another. Each race is represented by a number which can be converted to the actual race by looking at the document we prepared that determined which race each number represented.

Methodology/Models:
Regarding our methods, we first pre-processed and set up our data from the dataset. We then re coded part of the data such that each column had a 1 or 0 to indicate if the question it was asking pertained to the offender. We changed the data's coding to convert it from its original binary form to a multi-categorical one. We used a function that I created to combine the data from each row that related to the category column we were constructing into one. We assigned a number to each result, ranging from 0 to whichever many columns there were, where 0 was for a female and 1 was for a male, for example. When it comes to the data visualization and the models, it made it easier for us to utilize by appending them to create an array for each category value. We created our new column names, along with our X and y, based on these newly combined columns. We also developed a different function that used what we had learned, applied it to our training and testing data, and effectively recoded both sets of data. We then built a variable called col name to i that took the index of the column name and gave us the category of the column we were searching for so that each column name would properly function when we used it. I then started examining several models and applying them to my data. I considered and utilized three key models in this. I studied neural networks, random forests, and logistic regression.

Logistic Regression:
I began by studying logistic regression. I then fit the model to my training data. After that, I ran this model on my testing data using the.predict() method. The accuracy score() function was then used to gauge the predictions' accuracy using this model without correcting for bias based on these hypotheses. The accuracy was calculated at 67.74 percent.

I then built a function that would delete the columns pertaining to the primary protected characteristics in order to compute the accuracy without them after I calculated the accuracy before deleting any of the protected characteristics. In order to fit the logistic regression model to the new testing and training data without the protected characteristic, I was able to create this function for race, age, and gender as well. I then presented them in confusion matrices and used this to test the accuracy, precision, and f1 score for each one.

Random Forest Classifier:
Following my investigation into logistic regression, I considered employing a Random Forest Classifier model. First, I imported all the necessary statements. Next, I fitted my model to the training and testing sets of data I had available. Finally, I ran my model on the testing set of data to see whether the predictions would be more accurate than those using logistic regression. The Random Forest Classifier model had a comparable accuracy score of 65.94 percent, but I discovered that it was around 2% less accurate than the Logistic Regression model. The mean absolute error of the model was then calculated and evaluated, all without the protected features, and it turned out to be 34.06 percent.

Then, using the functions created to remove the protected characteristics, I went on to apply the same procedure used for the logistic regression model to the Random Forest Classifier model. I did this by calculating the accuracy score for each protected characteristic as well as the mean absolute error.

Neural Networks:
The neural networks model was the last model we used. We first imported all the libraries we intended to use for this model, as we had done with the previous two models, and then we created the variable "mlp" for our MLP classifier and fitted it to our training data. Then, without taking into account any of the influencing factors, I applied the model score function to the training and testing data. In contrast to the other two models, I only deleted one protected attribute from the neural network model in each test, thus I chose not to go into as much depth as I did for the other two models. The total accuracy score came out to be 65.48% which is similar to the Random Forest Classifier model, but is worse by the slightest amount.

Then, exactly as before, we performed the remove race and remove gender operations on the training and testing models, as well as the MLP classifier operation for the neural network model on both of them. We then conducted the operations to obtain the results and the confusion matrices.

Simple Model Metrics:
We developed our own methods to compute the accuracy, precision, recall, and f1 score for the simple model metrics that we used in our research. Then, we developed a function called subsetted statistics that needs the columns y true and y pred and outputs accuracy, precision, and recall for cases in which a response was not repeated as well as precision and recall in cases in which a response was. Then we built a variable called subset, which utilizes the unique function to discover distinct entries inside the list of the supplied column, produces a y true and a y pred for each, and prints out the accuracy, precision, and recall for each for both recidivated and not recidivated data. This was done for the gender and age columns as well as the race column to determine the accuracy, precision, and recall for each race represented in the dataset. Following that, the ROC curve, a straightforward model measure, was used. We calculated our y score by fitting it to our training data and used the decision function on our testing data to get our false positive rate and true positive rate, which served as our x and y axes. This was then plotted on a graph, and we were able to determine our AUC score using the graph and the findings.

```
Overall statistics for Logistic Regression Model:
    Accuracy: 0.6594269870609981
    Precision for not recidivated:  0.6711210096510765
    Precision for recidivated:  0.6401468788249693
    Recall for not recidivated:  0.7545909849749582
    Recall for recidivated:  0.5414078674948241
Statistics for Gender = 0 :
413
    Accuracy: 0.7360774818401937
    Precision for not recidivated:  0.7715231788079471
    Precision for recidivated:  0.6396396396396397
    Recall for not recidivated:  0.8534798534798534
    Recall for recidivated:  0.5071428571428571
Statistics for Gender = 1 :
1751
    Accuracy: 0.6413478012564249
    Precision for not recidivated:  0.6421052631578947
    Precision for recidivated:  0.6402266288951841
    Recall for not recidivated:  0.7254054054054054
    Recall for recidivated:  0.5472154963680388
```

Statistics for Gender column using Logistic Regression Model.

```
Overall statistics Random Forest Classifier:
    Accuracy: 0.6594269870609981
    Precision for not recidivated:  0.6711210096510765
    Precision for recidivated:  0.6401468788249693
    Recall for not recidivated:  0.7545909849749582
    Recall for recidivated:  0.5414078674948241
Statistics for Gender = 0 :
413
    Accuracy: 0.7360774818401937
    Precision for not recidivated:  0.7715231788079471
    Precision for recidivated:  0.6396396396396397
    Recall for not recidivated:  0.8534798534798534
    Recall for recidivated:  0.5071428571428571
Statistics for Gender = 1 :
1751
    Accuracy: 0.6413478012564249
    Precision for not recidivated:  0.6421052631578947
    Precision for recidivated:  0.6402266288951841
    Recall for not recidivated:  0.7254054054054054
    Recall for recidivated:  0.5472154963680388
```

Statistics for Gender column using Random Forest Classifier Model.

```
Overall statistics:
    Accuracy: 0.6709796672828097
    Precision for not recidivated:  0.6916403785488959
    Precision for recidivated:  0.6417410714285714
    Recall for not recidivated:  0.7320534223706177
    Recall for recidivated:  0.5952380952380952
Statistics for Gender = 0 :
413
    Accuracy: 0.7409200968523002
    Precision for not recidivated:  0.7677419354838709
    Precision for recidivated:  0.6601941747572816
    Recall for not recidivated:  0.8717948717948718
    Recall for recidivated:  0.4857142857142857
Statistics for Gender = 1 :
1751
    Accuracy: 0.6544831524842947
    Precision for not recidivated:  0.6670146137787056
    Precision for recidivated:  0.639344262295082
    Recall for not recidivated:  0.6908108108108109
    Recall for recidivated:  0.6138014527845036
```

Statistics for Gender column using Neural Networks Model.

```
Overall statistics:
    Accuracy: 0.6709796672828097
    Precision for not recidivated:  0.6916403785488959
    Precision for recidivated:  0.6417410714285714
    Recall for not recidivated:  0.7320534223706177
    Recall for recidivated:  0.5952380952380952
Statistics for Race = 0 :
114
    Accuracy: 0.6754385964912281
    Precision for not recidivated:  0.7283950617283951
    Precision for recidivated:  0.5454545454545454
    Recall for not recidivated:  0.7972972972972973
    Recall for recidivated:  0.45
Statistics for Race = 1 :
7
    Accuracy: 0.7142857142857143
    Precision for not recidivated:  1.0
    Precision for recidivated:  0.3333333333333333
    Recall for not recidivated:  0.6666666666666666
    Recall for recidivated:  1.0
Statistics for Race = 2 :
6
    Accuracy: 0.6666666666666666
    Precision for not recidivated:  0.5
    Precision for recidivated:  1.0
    Recall for not recidivated:  1.0
    Recall for recidivated:  0.5
Statistics for Race = 3 :
731
    Accuracy: 0.6922024623803009
    Precision for not recidivated:  0.7099811676082862
    Precision for recidivated:  0.645
    Recall for not recidivated:  0.8415178571428571
    Recall for recidivated:  0.4558303886925795
```

```
Statistics for Race = 4 :
192
    Accuracy: 0.7291666666666666
    Precision for not recidivated:  0.7591240875912408
    Precision for recidivated:  0.6545454545454545
    Recall for not recidivated:  0.8455284552845529
    Recall for recidivated:  0.5217391304347826
Statistics for Race = 5 :
1114
    Accuracy: 0.6463195691202872
    Precision for not recidivated:  0.6477495107632094
    Precision for recidivated:  0.6451077943615257
    Recall for not recidivated:  0.6073394495412844
    Recall for recidivated:  0.6836555360281195
```

Statistics for Race column using Logistic Regression Model.

```
Overall statistics:
    Accuracy: 0.6594269870609981
    Precision for not recidivated:  0.6711210096510765
    Precision for recidivated:  0.6401468788249693
    Recall for not recidivated:  0.7545909849749582
    Recall for recidivated:  0.5414078674948241
Statistics for Race = 0 :
114
    Accuracy: 0.6578947368421053
    Precision for not recidivated:  0.7011494252873564
    Precision for recidivated:  0.5185185185185185
    Recall for not recidivated:  0.8243243243243243
    Recall for recidivated:  0.35
Statistics for Race = 1 :
7
    Accuracy: 0.7142857142857143
    Precision for not recidivated:  1.0
    Precision for recidivated:  0.3333333333333333
    Recall for not recidivated:  0.6666666666666666
    Recall for recidivated:  1.0
Statistics for Race = 2 :
6
    Accuracy: 0.6666666666666666
    Precision for not recidivated:  0.5
    Precision for recidivated:  1.0
    Recall for not recidivated:  1.0
    Recall for recidivated:  0.5
Statistics for Race = 3 :
731
    Accuracy: 0.6908344733242134
    Precision for not recidivated:  0.6982142857142857
    Precision for recidivated:  0.6666666666666666
    Recall for not recidivated:  0.8727678571428571
    Recall for recidivated:  0.4028268551236749
```

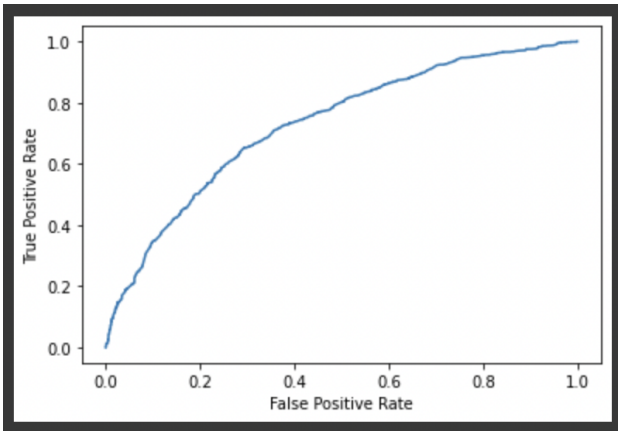
```
Statistics for Race = 4 :
192
    Accuracy: 0.7291666666666666
    Precision for not recidivated:  0.7320261437908496
    Precision for recidivated:  0.717948717948718
    Recall for not recidivated:  0.9105691056910569
    Recall for recidivated:  0.4057971014492754
Statistics for Race = 5 :
1114
    Accuracy: 0.6265709156193896
    Precision for not recidivated:  0.6196660482374768
    Precision for recidivated:  0.6330434782608696
    Recall for not recidivated:  0.6128440366972477
    Recall for recidivated:  0.6397188049209139
```

Statistics for Race column using Random Forest Classifier Model.

```
Overall statistics:
    Accuracy: 0.6709796672828097
    Precision for not recidivated:  0.6916403785488959
    Precision for recidivated:  0.6417410714285714
    Recall for not recidivated:  0.7320534223706177
    Recall for recidivated:  0.5952380952380952
Statistics for Race = 0 :
114
    Accuracy: 0.6754385964912281
    Precision for not recidivated:  0.7283950617283951
    Precision for recidivated:  0.5454545454545454
    Recall for not recidivated:  0.7972972972972973
    Recall for recidivated:  0.45
Statistics for Race = 1 :
7
    Accuracy: 0.7142857142857143
    Precision for not recidivated:  1.0
    Precision for recidivated:  0.3333333333333333
    Recall for not recidivated:  0.6666666666666666
    Recall for recidivated:  1.0
Statistics for Race = 2 :
6
    Accuracy: 0.6666666666666666
    Precision for not recidivated:  0.5
    Precision for recidivated:  1.0
    Recall for not recidivated:  1.0
    Recall for recidivated:  0.5
Statistics for Race = 3 :
731
    Accuracy: 0.6922024623803009
    Precision for not recidivated:  0.7099811676082862
    Precision for recidivated:  0.645
    Recall for not recidivated:  0.8415178571428571
    Recall for recidivated:  0.4558303886925795
```

```
Statistics for Race = 4 :
192
    Accuracy: 0.7291666666666666
    Precision for not recidivated:  0.7591240875912408
    Precision for recidivated:  0.6545454545454545
    Recall for not recidivated:  0.8455284552845529
    Recall for recidivated:  0.5217391304347826
Statistics for Race = 5 :
1114
    Accuracy: 0.6463195691202872
    Precision for not recidivated:  0.6477495107632094
    Precision for recidivated:  0.6451077943615257
    Recall for not recidivated:  0.6073394495412844
    Recall for recidivated:  0.6836555360281195
```

Statistics for Race column using Neural Network Model.




```
from sklearn.metrics import roc_auc_score

roc_auc_score(y_train, clf.predict_proba(X_train)[:, 1])
```

```
0.7071875309135371
```

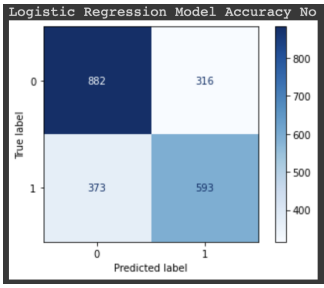
```
roc_auc_score(y_train, clf.decision_function(X_train))
```

```
0.7071875309135371
```

ROC AUC Score for the training data, along with the area of the graph.

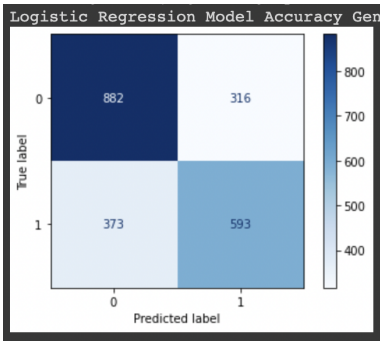
Results and Discussion:

Logistic Regression Model Accuracy No Race, Gender or Age: 68.16%



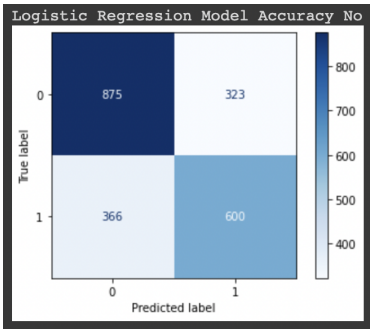
Random Forest Model Mean Absolute Error No Race, Gender or Age: 34.52%
Random Forest Model Accuracy Score No Race, Gender or Age: 65.48%

Logistic Regression Model Accuracy Gender or Age: 68.16%



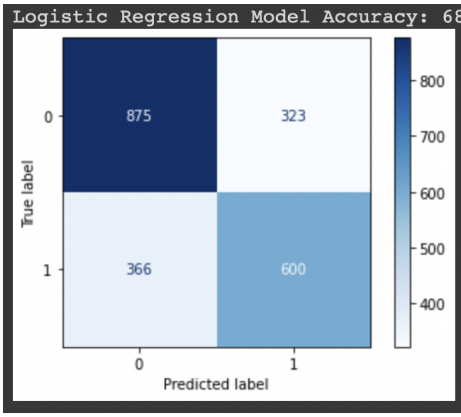
Random Forest Model Mean Absolute Error No Gender or Age: 34.52%
Random Forest Model Accuracy Score Gender or Age: 65.48%

Logistic Regression Model Accuracy No Race or Age: 68.16%



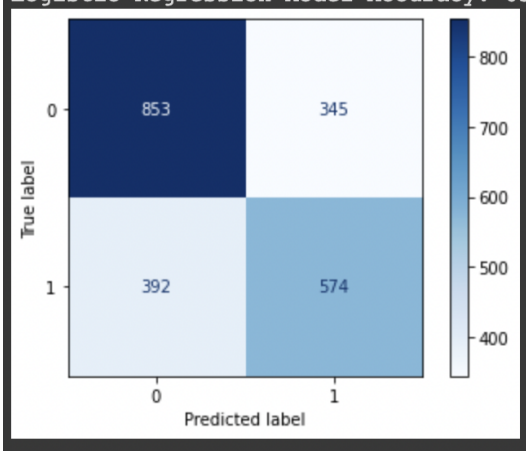
Random Forest Model Mean Absolute Error No Race or Age: 33.55%
Random Forest Model Accuracy Score No Race or Age: 66.45%

Logistic Regression Model Accuracy: 68.16%



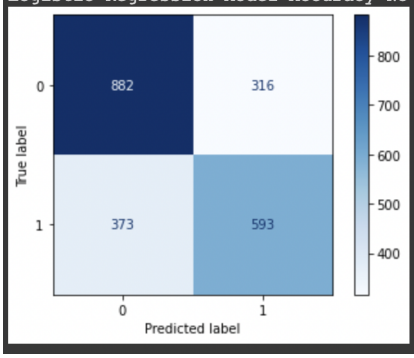
Random Forest Model Mean Absolute Error No Race or Gender: 34.52%
Random Forest Model Accuracy Score No Race or Gender: 65.48%

Logistic Regression Model Accuracy: 65.94%



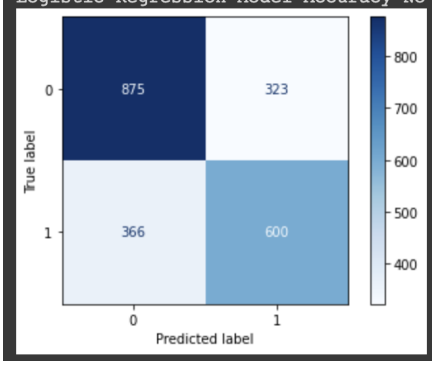
Random Forest Model Mean Absolute Error No Age: 33.55%
Random Forest Model Accuracy Score No Age: 66.45%

Logistic Regression Model Accuracy No Gender: 68.16%



Random Forest Model Mean Absolute Error No Gender: 34.52%
Random Forest Model Accuracy Score No Gender: 65.48%

Logistic Regression Model Accuracy No Race: 68.16%

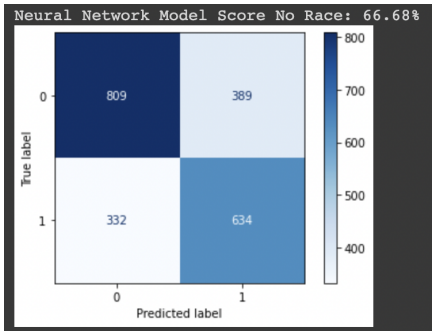


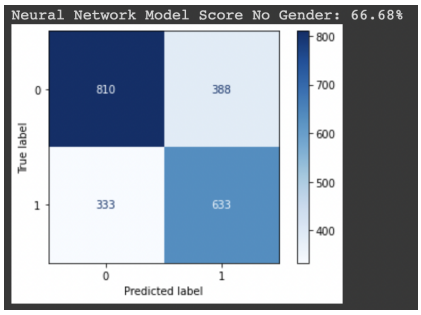
Random Forest Model Mean Absolute Error No Race: 33.55%
Random Forest Model Accuracy Score No Race: 66.45%

Neural Network Normal Training accuracy: 70.06%
Neural Network Normal Testing accuracy: 65.48%
Neural Network Model Accuracy Score Normal: 67.10%

Logistic Regression Model Normal Accuracy: 68.16%

```
Random Forest Model Normal Mean Absolute Error : 31.84%
Random Forest Model Normal Accuracy Score: 68.16%
```





Due to the wide range of findings made during this research, we will divide the comparison and data discussion into subsections after looking at all of the results provided.

Simple Model Metrics:

The simple model metrics for each protected attribute will be examined first. Every race is represented by a separate number. You can click on this link to get the paper containing this information. With 114 samples for Other, 7 samples for Asian, 6 samples for Native American, 731 samples for Caucasian, 192 samples for Hispanic, and 1114 samples for African American, the statistics for accuracy, precision, and recall for each different race show the difference in the variety of data used for each within the dataset. The differences between the samples utilized for each race may seem to indicate that the dataset is diverse, but it is really more skewed toward black men than white men. Even while they appear to be extremely excellent for some, if we look at the samples used for each one, we can see that the variation in the precision, recall, and accuracy between the samples is caused by the difference in the number of samples that are being utilized. The accuracy for Asians is approximately 86%, the precision for not recidivating is also approximately 86%, and the recall for recidivating is 1.0; however, the accuracy for Native Americans is significantly lower at approximately 66%, the precision for not recidivating is lower at approximately 65%, and the recall for not recidivating is higher at approximately 96%. In addition to the different sample sizes (7 Asians and 731 Native Americans), these discrepancies are also the result of all previous accounts in which judgments were made using protected variables that cannot be changed. Native Americans and African Americans both had similar accuracy levels, with Native Americans having an accuracy of 66 percent and African Americans having an accuracy of 65 percent, when the precision and memory for recidivated and not are compared. When accuracy and recall are compared, the precisions are somewhat more comparable than the recall. Native Americans have an accuracy for not recidivating of approximately 65 percent, whereas African Americans have a precision for not recidivating of around 62 percent, which is obviously similar but slightly worse at the same time. Native Americans have a remember rate of 96% when it comes to not recidivating, compared to African Americans who have a recall rate of 71%. This difference is significant, and it may have been seen because of earlier convictions that have been utilized or because these protected qualities may have clouded one's judgment. The same logic applies when we consider recall and accuracy for non-recidivism. Native Americans have a precision of around 75% when it comes to recidivism, whereas African Americans have a precision of about 66%. Native Americans have a memory rate of 17%, whereas African Americans have a remember rate of 59%. This research may suggest that, when compared to African Americans, who are more likely to be considered as likely to recidivate, it is more probable that a Native American will

be regarded as not recidivating. We observed several variances, particularly within these two races, hence the conclusions drawn for this are solely dependent and concentrated on 2 of the 6 races. The recall for those who recidivate is so high comparatively for Black individuals (58% while it's under 20% for a lot of the other categories). In contrast, recall for those who do not recidivate is very high for white individuals (96% while it's around 70% for Black individuals). Thus we suggested that we were finding a similarity to the ProPublica expose where Black individuals seem to have higher accuracy rate for recidivated individuals while white individuals have a higher accuracy rate for not recidivated individuals, displaying your model does exhibit some bias in how it assigns its predictions.

Confusion Matrices & Removal of Protected Characteristics:
Unexpected findings were discovered while examining the various models employed and how their accuracy varies as a result of adjustments made, such as dropping columns. First, when examining the accuracy of the Logistic Regression, Random Forest Classifier, and Neural Networks models, we first examined the accuracy without removing any of the protected characteristics, and the accuracies were similar across all models. The accuracy of the Logistic Regression Model was 67.74 percent, the accuracy of the Random Forest Classifier Model was 65.94 percent, and the accuracy of the Neural Networks Model was 65.48 percent. This precision is already adequate but not outstanding. During our background investigation, we discovered that the best model was Logistic Regression. One of the publications had discovered that their Random Forest model provided them with superior outcomes. Following that, we began to eliminate each of the three protected features that we believed to be the root of this injustice. At first, we just eliminated one of the traits at a time. With a mean absolute error of 33.55 percent, the accuracy of the Random Forest Model was 66.45 percent and that of the Logistic regression was 68.16 percent and the Neural Networks Model was 67% when race was removed. When gender was removed, the accuracy of the Random Forest Model was 65.48 percent, the accuracy of Logistic regression was 68.16 percent, the accuracy for Neural Networks was 67% and the mean absolute error was 34.52 percent. When age was removed, the accuracy of the Random Forest Model was 66.45 percent, while the accuracy of Logistic regression was 65.94 percent, with a mean absolute error of 33.55 percent. This suggests that although age may be considered a protected factor, it may not have as much of an impact as gender or race, which are considered to be the more significant protected qualities. In contrast, when age is removed, there is more of an improvement being made. Then, we eliminated two traits at a time. When race and gender are taken into account, the accuracy of the Random Forest classifier is 65.48 percent, while the accuracy of Logistic regression is 68.16 percent, with a mean absolute error of 34.52 percent. When race and age are taken into account, the accuracy of the Random Forest classifier is 66.45 percent, while the accuracy of Logistic regression is 68.16 percent, with a mean absolute error of 33.55 percent. Age and gender were removed, and the accuracy of the Random Forest classifier was 65.48 percent with a mean absolute error of 34.52 percent. The accuracy of the Logistic regression was 68.16 percent. This suggests that the findings are slightly worse when the two key protective characteristics—race and age—are removed than when they are removed together. Gender may be significant in some circumstances while examining a case, thus that protected attribute may occasionally be required. The accuracy is somewhat worsened when gender is removed. Then, we simultaneously deleted all three traits. When age, gender, and race were taken into account, the accuracy of the Random Forest classifier model was 65.48 percent, with an absolute mean error of 34.52 percent, while the accuracy of the Logistic regression model was 68.16 percent. It is evident from this that when these qualities are removed, the accuracy is marginally poorer than when we didn't remove any factors at all. When analyzing the entire body of data, it becomes clear that even when these protected characteristics are removed, the accuracy remains consistently frequent and similar, suggesting that having these protected characteristics may play a role in if individuals recidivate again. . There is a caution that

should be mentioned, which is that, in our opinion, when encoding real world biases, if our algorithm is picking up on them it is likely to perpetuate them, even if the current dataset breakdown reflects a system that is unfair or biased.

Conclusion:
In light of the analysis of all the available information, it may be said that the research issue is still unclear. However, based on the results shown and the information found, our work suggests that in the particular models we examined that the accuracy of determining how likely an offender is to recidivate does not especially improve if certain protected characteristics are retained. We hadn't managed to look at other hyperparameters and other models, which would help give us a better idea of our results. It is also possible to draw the conclusion that there is some injustice in this regard, as was demonstrated when examining the accuracy for each group of protected features. Additionally, the dataset we utilized included data that may be viewed as biased due to the disparity in sample sizes employed for each race (for example, 1114 African Americans were included compared to 7 Asians), which may affect the findings of this research topic. Further, we saw that there were disparities in what kinds of false predictions our models made for different races, reflecting the results of the ProPublica expose. The results may be impacted in various ways, therefore there is currently no conclusive response to this topic. However, our research indicates that protected qualities do not significantly alter assessments of an offender's likelihood of recidivism, and in some circumstances worsen them.

Acknowledgements:
I want to thank my boss for working with me throughout the process, helping me fix the code's faults, and helping me create my research algorithm. I also want to express my gratitude to the software package InspiritAI for enabling me to conduct this study.

References/Bibliography:
(Angwin, Larson, Kirchner and Mattu, 2016) - Angwin, J., Larson, J., Kirchner, L. and Mattu, S., 2016. *How We Analyzed the COMPAS Recidivism Algorithm*. [online] ProPublica. Available at: <https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm> [Accessed 9 August 2022].

(Hao and Stray, 2019) - Hao, K. and Stray, J., 2019. *Can you make AI fairer than a judge? Play our courtroom algorithm game*. [online] MIT Technology Review. Available at: <https://www.technologyreview.com/2019/10/17/75285/ai-fairer-than-judge-criminal-risk-assessment-algorithm/> [Accessed 9 August 2022].

(Mehta, Shah, Patel and Kanani, 2020) - Mehta, H., Shah, S., Patel, N. and Kanani, P., 2020. *Classification of Criminal Recidivism Using Machine Learning Techniques*, 29, pp.5110 - 5122.