Recidivism Final Project Report

Andi Zhao, Helen Wu, Ryan Sun

ABSTRACT

In a situation where algorithms dictate the future of humanity, one must proceed with extreme caution. For example, *COMPAS*, a software that is employed by the legal system to predict the recidivism rate of criminals, came under fire for neglecting the aspect of fairness in 2016. The ProPublica organization discovered racial bias within the predicted results, indicating a lack of thought and consideration of ethics while creating the algorithm. Using a dataset from the Iowa state government correctional center website, our final project focuses on this subject matter by attempting to predict recidivism with and without eliminating protected features from the training set to see if it has any effect on predicting recidivism. Pre-processing was required to eliminate unnecessary features and missing information. This report will detail the process in which we gathered and processed the data, as well as the conclusions that we drew from comparing our different models.

1 MOTIVATION AND OBJECTIVE

In recent years, the traditional justice system has been accused of being biased to minority populations. In an attempt to resolve this issue, researchers developed *COMPAS* (Correctional Offender Management Profiling for Alternative Sanctions), a tool used to circumvent the need for human judgement. The reason that it came under fire was because it was yielding biased results where "blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend"[1]. In addition, privacy and fairness have been at the forefront of modern artificial intelligence research. Some researchers have attempted to train models without protected attributes to avoid lawsuits. We wondered if eliminating protected attributes from the training set actually has any effects on the accuracy and precision of the model. If the protected attribute has no effect on the accuracy of the model, it would mean that either race has no effect on the outcome of the model or that race was predicted. On the other hand, if race does have an effect, it would mean that the researchers were correct to hide the protected attribute.

Our intention is to use Iowa's recidivism dataset to train a model with and without race as one of its features to determine if race is a factor in determining recidivism. We also attempt to predict race from recidivism data to see whether or not the results are significant enough to associate race and recidivism. Our plan was to train two models, one with logistic regression and the other with decision trees, to compare accuracy, and other statistics. The goal of this project is to produce models that predict with similar accuracies to that of *COMPAS*, as well as to uncover the relationship between using race as a feature and the corresponding results for predicting recidivism risk scores.

2 ETHICAL ISSUES

The ProPublica study showcased the dangers of algorithmic decision making without concern for ethics, namely race, which can be detrimental to a true understanding and accurate modeling of recidivism. Additionally, the institutional biases within the *COMPAS* algorithm directly affects our justice system because *COMPAS* risk scores are actually considered by judges during sentencing in certain US states. The chart below shows the flaws and clear biases of blindly using the *COMPAS* assessment to identify risk scores for an individual during sentencing.

Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Figure 1: ProPublica's discovery

In our study, we are primarily concerned about whether or not using race as a feature will affect

the prediction of recidivism for an individual. By including and excluding race as an attribute, we are able to create different models and compare their results, thus allowing us to determine whether or not race was a significant factor in the predictions. We also tried the opposite, using recidivism to predict race, to see if any ethical concerns arise.

3 DOMAIN AND DATASET

Initially, we wanted to use ProPublica's released data set as our main source of data. However, after analyzing the CSV file, we realized it was too convoluted to understand. Many of the column names were not self-explanatory and the documentation for each feature was lacking. Furthermore, there were duplicate information for some of the individuals, and cleaning and analyzing the data set would have taken far too long. After speaking to our TA about our project, we decided to utilize another data set released by Iowa state's government correctional center facility. Not only were these feature names straightforward and easy to understand, the data was cleaner and more recent.

The data set contained recidivism information for 26,020 individuals from 2007-2013 in the state of Iowa. In total, there were 17 columns, with the "Return to Prison" column indicating whether the individual recidivated within the last 3 years. Out of the 17 attributes, age, race, and sex were the only 3 protected attributes in the data set. To protect the identities even further, each individuals age at release fell into a 5 categories; no actual ages were revealed. Strangely, the race category did not contain a unique Hispanic value as each of the other race groups were either Hispanic or Non-Hispanic (this can be seen in the graph below). This phenomenon in the data set is unusual as according to the 2010 US census, the Hispanic or Latino population was around 6% of the total population, making it the second highest ethnic group population in the state.[2]

After reviewing the data set, we decided to choose 5 non-protected attributes we thought were beneficial and relevant to creating our models. These 5 features were: Release Type, Age At Release, Offense Classification, Offense Type, and Target Population. Since we also wanted to test whether or not race affected the prediction results, we also added the race attribute as one of our features for two of our four models. While the former 5 features were relatively clean, the latter race feature was a little more difficult. Incomplete data (i.e., 'White -', 'Black -', 'NA -') littered the race column, and since we were not able to differentiate which group these incomplete values belonged to, we removed them from our data set. In addition, we replaced NaN with -1 and proceeded to remove any rows with missing values to avoid future complications of fitting models. This brought our total data set size down to 5,195.

Before we removed any rows, we wanted to see the original data set's ethnic group statistics and how it could impact our results. In Figure 2, the number of White and Non-Hispanics greatly outnumber the rest of the ethnicities. As we will see later, the heavily skewed data impacted the predictions of the four models we created. The chart below is a visual representation of the original dataset grouped by ethnicity and recidivism count.

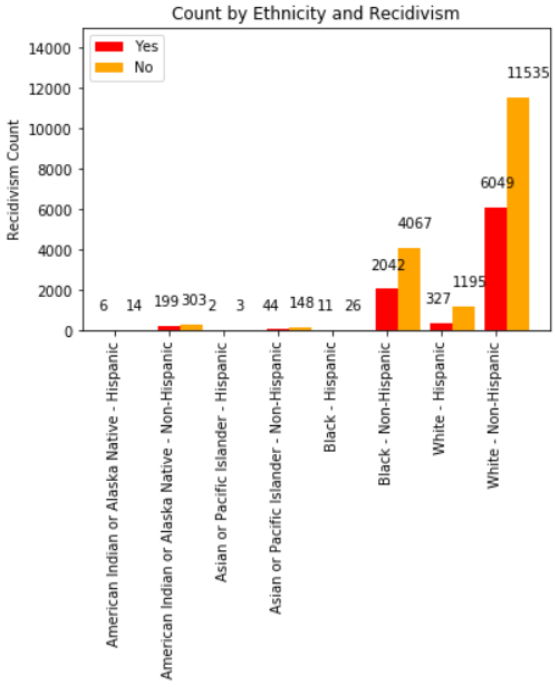

Figure 2: Data Distribution

4 MODELS AND ALGORITHMS

4.1 Decision Trees

In this project, we primarily used sklearn's decision tree as well as logistic regression packages to model the data. To keep things modularized, we created multiple models to predict both race and recidivism. When predicting for recidivism, we used both decision tree as well as logistic regression models to verify and compare the results. Models included and excluded the race feautre to check if race impacted recidivism predictions. When predicting for race, we used a decision tree with the recidivism feature.

Before we began creating the decision tree model, we first had to find which depth would minimize error. As seen in Figure 3, we discovered that in order to minimize error (based off of accuracy), the longest path from root to leaf went no more than 6 nodes. Therefore, all of our decision tree models had a depth of 6.

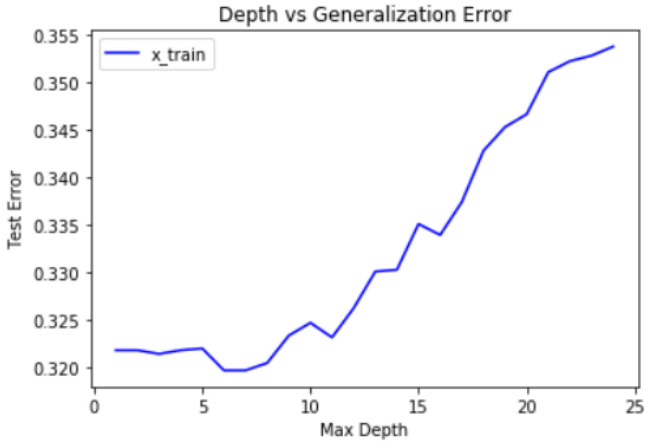

Figure 3: Decision Tree Depth Analysis

After the decision tree depth analysis, we realized that many of the features that we selected for this project were mainly categorical features, such as "Target Population" or "Offense Classification." This required the transformation of each unique value of a categorical columns into separate 0-or-1 columns before being appended back into the dataset. The model that included race as a feature would introduce many additional columns in the dataset, one for each unique value in the original column. After the transformation of the dataset into binarized values, the data is split into the testing and training sets before creating the model to predict for recidivism. To decide the depth parameter for the decision trees, we used accuracy as a measure of optimization for deciding the tree depth to use.

In an attempt to predict race from recidivism features, the features that were previously removed from the data set were used. The new dataset with information about recidivism also had mostly categorical features, so the same process as before was repeated in order to binarize the features for a decision tree model. This model would predict on the 'Race' labels, and took the same approach as the decision tree models detailed above.

4.2 Logistic Regression

Originally, we planned to used sklearn's logistic regression as a second potential model. Due to the nature of a logistic regression, the inputs must be categorical, so we converted all features into category type data, which allowed us to assign a number to all entries of our feature vectors. We decided to split the data into 20/80 train and test sets respectively to keep consistency between models and trained a model excluding race as a feature. Similar to the decision tree models, we then trained a separate logistic model with the same train test split, but included race as a sixth feature. The approach of the logistic regression was very similar to the decision tree model approach as the logistic model was intended to verify the results of the decision tree models.

5 RESULTS AND ANALYSIS

To analyze the data of the decision trees, we chose to look at the accuracy, precision, and recall of the results. In the model where the race attribute was excluded as a feature, we achieved an accuracy of 66.7%, similar to that of COMPAS' (63-67% depending on race). The exact numbers our model had predicted was 213 cases of recidivism and 4,982 non-recidivism cases. The labels for the testing data indicate 1,729 cases of recidivism and 3,466 non-recidivism cases. The next step was to create a model that predicted with race as a feature, and the results were similar: 225 predictions of "Yes"

for recidivism and 4,970 predictions of “No.” The difference was around a dozen more for “Yes” when race was used as a feature to predict for recidivism. The charts below are the results graphed with matplotlib, and they are identical.

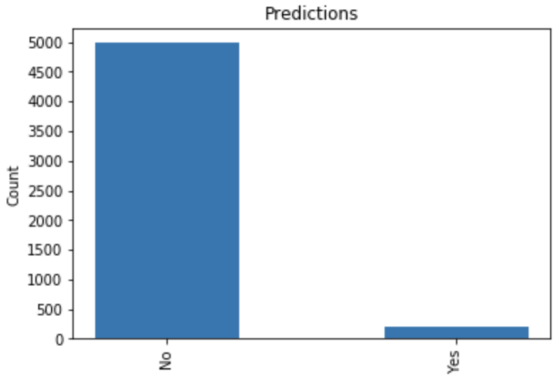

Figure 4: Decision Tree: Without Race

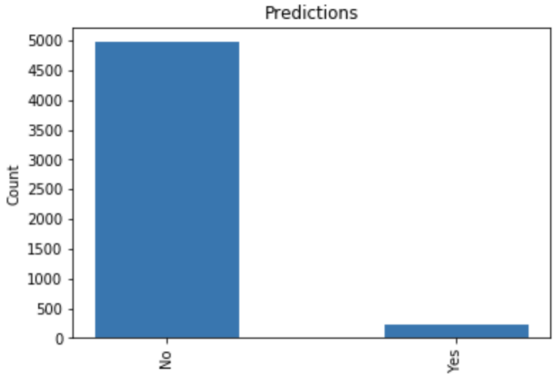

Figure 5: Decision Tree: With Race

From a speculative standpoint, we brainstormed 2 possibilities. The first possibility is that race does not affect the outcome of the model, since the only difference in the models were the features they were trained on. However, the more plausible explanation for our findings is the second possibility: there exists some combination of proxy variables within our 5 features that could predict race. In order to ascertain our beliefs, we created yet another model to predict race from recidivism features. The model predicted the following:

Race (Condensed)	Predictions	Actual
Indian - Hispanic	0	5
Indian - Non-Hispanic	0	94
Asian - Hispanic	0	1
Asian - Non-Hispanic	3	35
Black - Hispanic	0	9
Black - Non-Hispanic	75	1244
White - Hispanic	23	302
White - Non-Hispanic	5094	3505

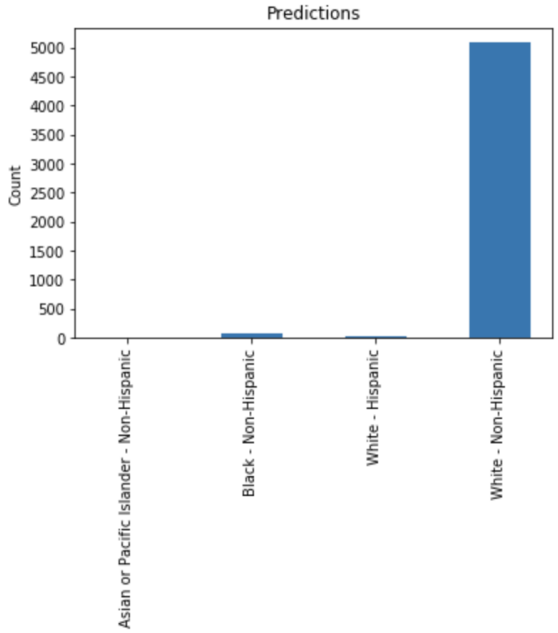

Figure 6: Predicting Race

This data still clocks in at 68% accuracy, similar to the accuracy of the model predicting recidivism. Unfortunately, although it produces similarly skewed results to the original data set, we quickly realized the accuracy was decently high only due to the multitude of "Yes" for ’White - Non-Hispanic’ group. This model is predicting mainly ’White - Non-Hispanic’ for almost every entry in the test data, with only around 100 predictions combined from the other ethnic groups. In fact, this situation is similar to the accuracy example of predicting every case to be false for a 1% likely disease and obtaining a 99% accuracy score. In the end, we could not definitively prove whether or not proxy variables were in play when producing the models.

Perhaps optimizing the decision tree for accuracy may not have been the best choice, but it could also be simply due to the fact that our data represented the race groups differently as well. These decision tree results tell us that there are some intrinsic properties of this dataset that skews our results. Perhaps we need more features and data representing all ethnic groups equally to train and create a more accurate model.

The logistic model confirmed the results of the decision tree model. The regression model had an accuracy of 67% and a similar precision at 68%. In addition, the predicted individual counts between models was very similar, suggesting that there were no errors in the training process.

While this could be a case of biased source data, the results clearly state that hiding race did not affect the predictions of the models. In the future, we plan to confirm the conclusion by using data from the actual COMPAS dataset. Note: This conclusion does not state that race did not play a role in determining recidivism, it simply states that eliminating a protected attribute such as race as a feature does not mean the protected attribute did not play a role in the predictions.

6 CONTRIBUTION

Ryan Sun created the decision tree models for predicting recidivism with and without race as a feature. He also created the model for predicting race from including recidivism features. Along with the group, he took part in the analysis of the results and created charts for the poster/video presentation. He also took part in authoring multiple sections of this report.

Huanlei Wu did most of the pre-processing and pre-processing visualizations. At first, she considered adding neural networks and random forest classifiers as models, but ultimately decided against it. However, the code is still in the Jupyter notebook. She also helped Andi Zhao with some of the visualizations for the logistic regression model at the end of the project, mainly the two recidivism prediction graphs. In addition, she also took up the task of cleaning up and beautifying the group's code so it was more readable and easier to understand. Along with the whole group, she helped analyze the results, create the presentation poster, as well as write and proofread the final report. Her focus of the report was mainly on the "Domain and Dataset" portion.

Andi Zhao worked on logistic regression as a second model to compare with the decision tree models. Andi Created two models, one included race as a feature and one without. He outlined the goal of the project and provided a roadmap on how to approach the goal. He provided the template for the poster and wrote the script for the video presentation. During the video presentation, he outlined our motivation and introduced the project.

7 FUTURE WORK

An interesting follow-up for this project would be to create models for the original COMPAS dataset and compare the results to that of this project's. Does the difference in race representation between the large datasets in the data skew the results? Or are there further confounding factors present? Comparing the results to our current models could potentially provide further insight into the influence of race as a feature in machine learning models.

Taking it a step further, we could also look at *many* data sets and see how the protected attributes, especially race, affects recidivism predictions. We could look at more than just the basic measurable statistics of these results and explore many other possible connections as well. Further inquiries could possibly be made to pin down the true factor of race and ethnicity in these machine learning projects and algorithms.

Finally, although a little out of our abilities, understanding which features played how much in determining the outcome would also be a viable future goal. For our project, we would specifically look at which features contribtuted how much to predicting race. Knowing the weight of each feature would afford us a better understanding and analysis of our results.

REFERENCES

[1] Angwin, Julia; Larson, Jeff (2016-05-23). "Machine Bias." ProPublica. Retrieved 2019-11-21.

[2] "U.S. Census Bureau QuickFacts: Iowa." Census Bureau QuickFacts, www.census.gov/quickfacts/-fact/table/IA/POP010210.