Abstract

This paper discusses the possible ethical implications of models being trained with the Census Income Dataset, which concludes whether or not a person's income is greater than $50,000 per year. We'll analyze the overall performance of our models, specifically the number of false negatives our model outputs for different subgroups in our dataset. Additionally, we'll touch on different aspects of fairness in machine learning that offer an explanation as to why our algorithm falters when isolating certain attributes based on various thresholds to the respective data. We begin by presenting our motivations for this paper and what we hope to convey to its readers. We go on to explain our dataset in detail and provide an in depth analysis of our algorithm, discussing important areas relating to fairness, such as biases inherently present in the dataset and possible ones from our research that was analyzed in retrospect.

Introduction and Motivation

Income prediction can be a powerful tool for businesses. For example, a company may choose to expose customers who earn higher than $50,000/year to more premium products than their counterparts. Real Estate Agencies may follow similar behavior when deciding which selection of houses to show to customers. Loan agencies may be more inclined to give loans to people who make more than $50,000. If all these businesses were making their decision based on this dataset, then there could be issues down the road. Ideally all these companies want to capitalize on their profit margins and that occurs with individuals who can offer them the most money(people who make more than $50,000). With that in mind companies want to maximize the number of predictions for those who make >50,000(True Positives) and number of predictions for those who make <$50,000(True Negatives). However, when incorrect predictions occur, individuals and businesses face real consequences. In the event they want to be risk averse and prefer incorrectly predicted cases for those who make <=$50,000. This may result in a loss of potential customers but it also prevents the company from great losses(False Negative). The company would want to minimize the number of people the model predicts >$50,000 but in reality they make <=$50,000, as this would hurt the companies the most(False Positive). In an ideal world, our models would be 100% accurate. We will prioritize analyzing the false negatives across several significant features, which we will identify through threshold testing and features we believe have inherent biases. In all of this we focus on two main points of view: the company/business and the individual; these play a role in the type of biases either may choose to be implementing.

Motivation and Goal

Discuss what fairness is

Use charts and stuff, like history of fairness in ML.

Fairness has been given much more importance in machine learning over the past several years. Why? It has a direct impact on how successful our algorithms are. Tasks/decisions that directly affect humans are increasingly automated by machine learning, thus fairness in ML and human benefit are becoming rapidly intertwined.

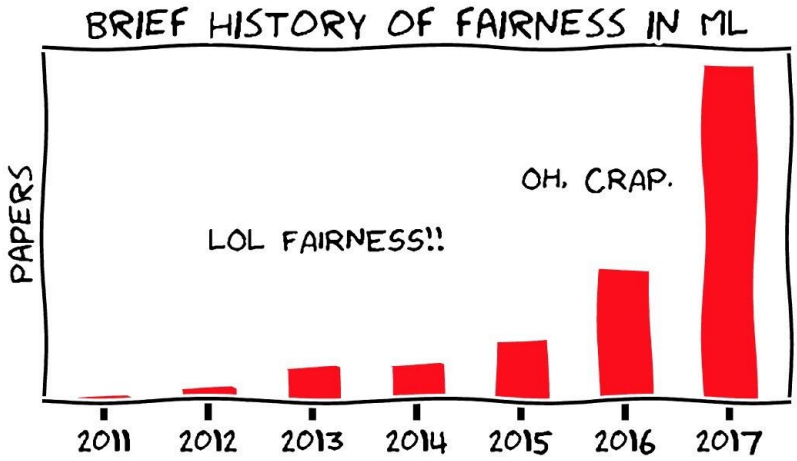


Figure: Rising number of publications on fairness in ML.

Next follows a few examples of predictive models outputting unfair results.

In 2016, a study conducted on COMPAS, a tool used to predict recidivism, showed that the algorithm has a much higher false positive rate for African Americans. These predictions have a direct impact on a defendant's incarceration length, so a wrong prediction can have terrible consequences.

	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%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. (Source: ProPublica analysis of data from Broward County, Fla.)

Figure: COMPAS false positive rates for White/African Americans.

Additionally, a job platform called XING was found to rank female candidates lower than their less qualified, male counterparts.

Search query	Work experience	Education experience	Profile views	Candidate	Xing ranking
Brand Strategist	146	57	12992	male	1
Brand Strategist	327	0	4715	female	2
Brand Strategist	502	74	6978	male	3
Brand Strategist	444	56	1504	female	4
Brand Strategist	139	25	63	male	5
Brand Strategist	110	65	3479	female	6
Brand Strategist	12	73	846	male	7
Brand Strategist	99	41	3019	male	8
Brand Strategist	42	51	1359	female	9
Brand Strategist	220	102	17186	female	10

TABLE II: Top k results on www.xing.com (Jan 2017) for the job serach query "Brand Strategist".

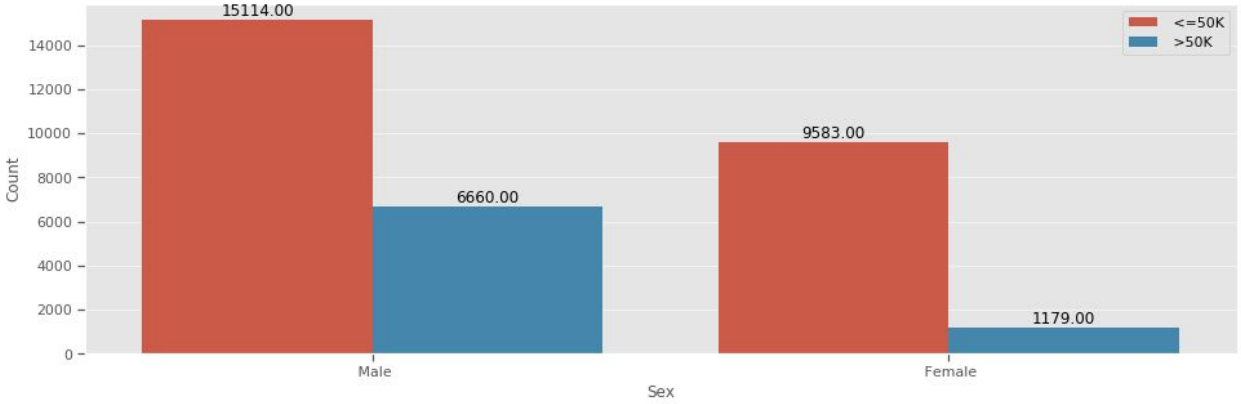
Figure: Top 10 candidates for 'Brand Strategist'

Our goal is to present rudimentary models that showcase biases in the data and perhaps what causes these biases. We present our findings so readers understand how our model gives very different results when isolating certain attributes based on various thresholds we choose to be significant. Additionally, we hope to show our readers methods to identify possible causes of this unfairness.

Ethical Issues

We are investigating the issue of fairness in predictive models being trained with data from the Census Income Dataset. Many of the problems appear to be underlying, such as how our data was gathered in the first place.

When initially analyzing our model on subgroups divided by the gender attribute, we noticed that there were more than twice as many instances of male candidates than female candidates. Furthermore, there are almost six times as many instances of male candidates who earn over $50,000 than female candidates. This uneven distribution of entries can be partly attributed to **reporting bias,** which occurs when the frequency of events, properties, and/or outcomes captured in a data set does not accurately reflect their real-world frequency. It could additionally be attributed to the historical context given the Census was taken in 1996 and historically men made more than women so this could also attribute to the stark differences in the context of this feature. Thus, given the difference in sample sizes for each group and historical context, the error rates may not be an accurate representation of the model's accuracy when pertaining to the current time frame.



Number of instances by gender

Additionally, we noticed that the dataset predominantly contained information instances where individuals made less than $50,000/year. Based on the dataset about 75% of individuals made <=$50,000. We'd like to point out that this dataset is not an accurate reflection of individuals' incomes today. In 1996, the median household income was $35,492. However in 2016, the U.S. Census Bureau reported that real median household income was $59,039, a more than $20,000 increase from 1996.

Furthermore, we observed a possible data skew in the set. There is no feature to help us identify the area in which each individual in the dataset resides. If the entries in the dataset come from individuals that are geographically close to each other, our data may not accurately reflect the United States population as a whole. This could be due to a result of a privacy issue. In order to proceed further we assume the data is IID.

There is a possible sample size disparity. If the training data from a minority group is much less than the majority, the model is less likely to perform perfectly on the minority group. We noticed for the native country attribute, a majority of our dataset were from the United States (91.388%) followed by Mexico (1.964%) and decided to feature engineer an "Others" category containing the rest of the countries.

Domain/Dataset

The prediction task of the Census Income Dataset is binary. That is, it outputs whether or not a person makes over $50,000/year. Below is a listing of the various attributes and the values each can take on.

Age: a continuous numerical value.
Workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
Fnlwgt: a continuous numerical value.

education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
Education-num: a continuous numerical value.
Marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
Occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
sex: Female, Male.
capital-gain: a continuous numerical value.
capital-loss: a continuous numerical value.
hours-per-week: a continuous numerical value.
native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.
Income: >=50k, <50k

Data Preprocessing (subsection of Domain/Dataset)

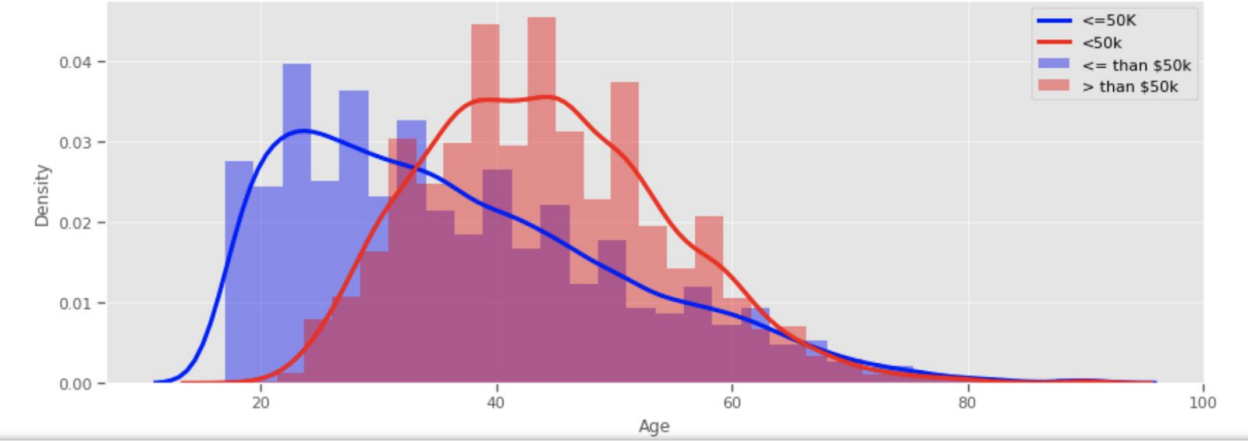
Based on the dataset population, 75.91% people were making less than or equal to $50,000/year and 24.09% people were making more than $50,000/year. We feel the dataset accurately reflects income distribution from the time it was collected, we did not perform any data normalization.

Our dataset includes both numerical and categorical data. We analyzed numerical data features with distribution plots between the two groups(<=50k and >50k), displaying histograms and a Kernel Density Estimate plot. We also used box plots for numerical data to display the distribution and account for any significant outliers. For categorical features, we displayed our data using count plots.

Based on our data we observed that
KDE (Numerical Data)

- Age: ranged from 17-90, and ages >30 classified to make >50k
- Fnlwgt: disregard this feature, data for both groups is at same density
- Education Num: plausible significant feature
- Capital Gain: doesn’t appear to be a significant feature, too skewed
- Capital Loss: doesn’t appear to be a significant feature, too skewed

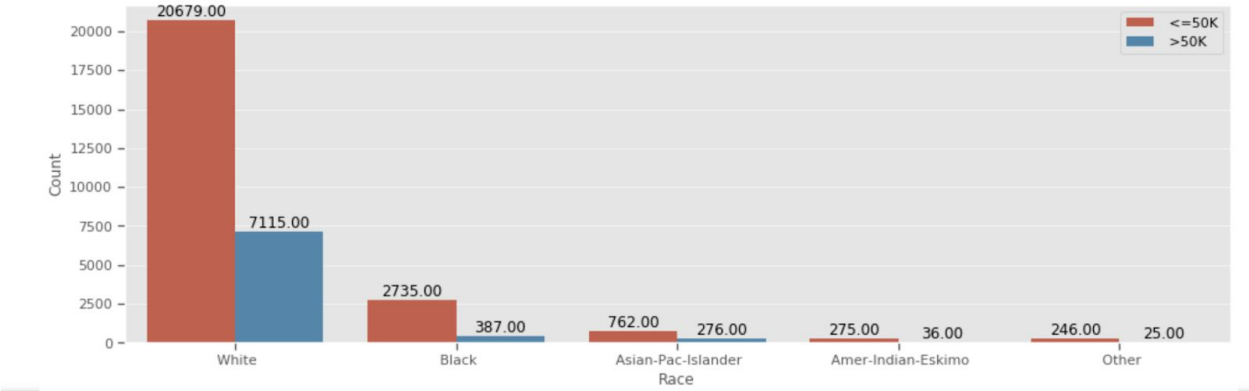
- Hours per week: ~40% can say that if they work 40 hours/week earn <=50k, 12% who work 40 hours/week earn > 50k



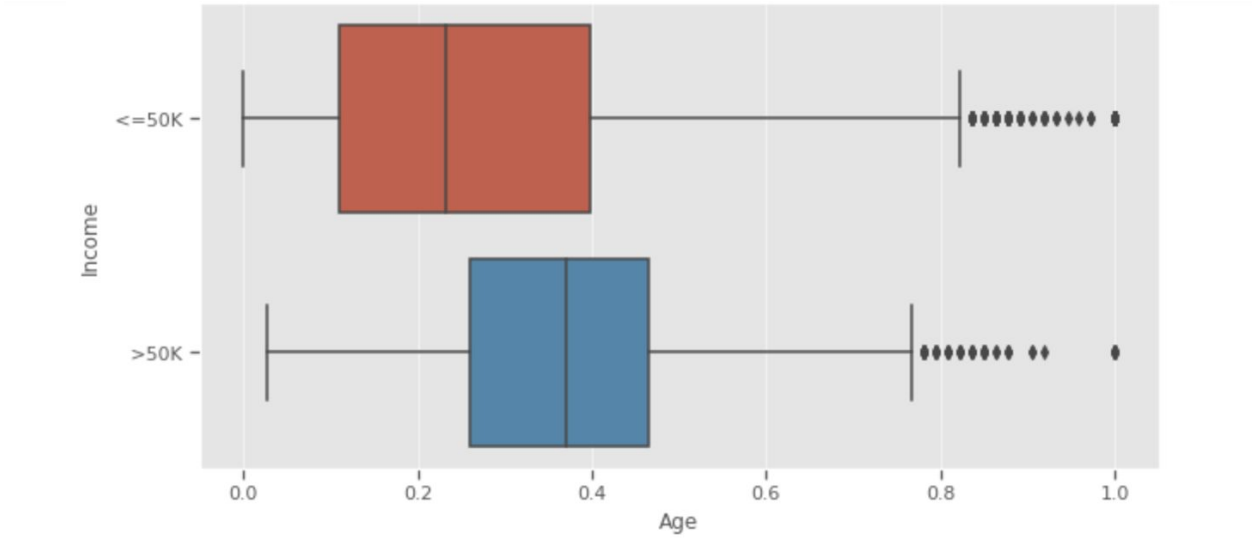
KDE distribution plot of Age attribute, overall we noticed individuals older than 30 make more than $50,000.

COUNT PLOT(Categorical Data) -

- Workclass: passable significant feature
- Education: numerically identical (goes with Education Num numerical feature)
- Marital status: plausible significant feature
- Occupation: Various categories, plausible significant feature
- Relationship: possible correlation w/ marital status
- Race:
    - 74.4% of whites make <= $50,000
    - 87.6% of blacks make <= $50,000
- Gender:
    - 69.4% of males make <= $50,000
    - 89% of females make <= $50,000
- Native Country: so many categories, note top 3 countries



Count plot of Race attribute, majority of our dataset is white and percentage of black making less than $50,000 is greater than whites.



Box plot of Age attribute, shows distribution in 4 quartiles and notes any significant outliers.

Initially we constructed our **base dataframe** where we dropped the missing values and manually encoded the categorical features.

Then we constructed two other dataframes called One Hot Encoding(OHE) and Binary Encoding(BE) where we performed some feature engineering on the raw dataframe.

Feature Engineering (subsection of Domain/Dataset)

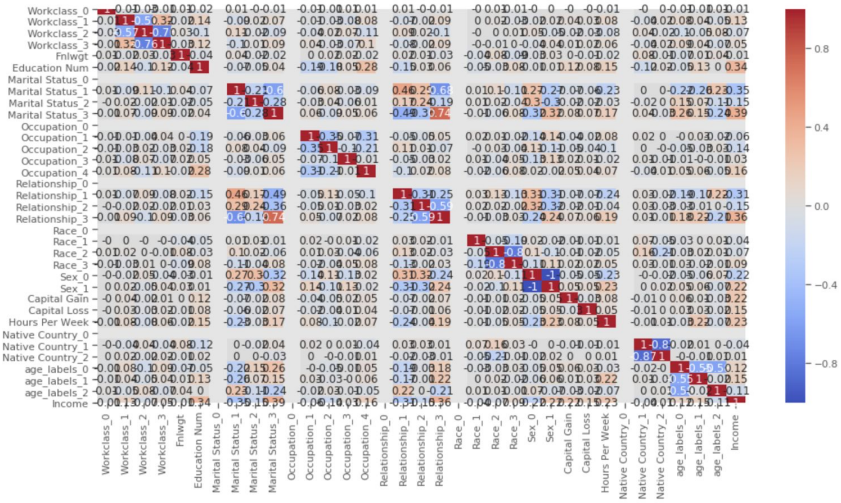
After analyzing our data's features, we encoded the categorical data differently since it was nominal data. This would increase our chances of achieving higher accuracy for our two main algorithms we intended on using, logistic regression and decision trees. We chose One-Hot-Encode one of them since it favors Logistic Regression models and Binary-Encode the other for our decision tree algorithm as it results in a smaller cardinality of features.

In our initial model, there were 2,399 rows that were missing values, which were dropped from the dataset. However for the dataframes mentioned above, we imputed missing/unknown values instead of dropping them since they were not significant enough to be dropped. Features containing missing/unknown values in our dataset include: workclass (5.643%), occupation (5.664%), and native country (1.789%). We replaced the missing values with the mode of the respective feature's column. Additionally, we bucketed age into quartiles: Young (17-28), Middle-Aged(29-37), Old (38-48), Retired (49-90) since this would aid both models in decision making. For the native country attribute, we featured engineered an "Others" category because all countries other than the United States and Mexico were insignificant in data representation. For all our numerical values, we scaled the data points to a range of values (0-1) as it prevents

the model from prioritizing one feature over another and the numbers ranges varied drastically. Finally we noticed that there were some entries which were duplicate values and decided to drop the duplicate values. This could be attributed to some sort of error while collecting the data.

After these changes, we decided to create a heat map to observe initial correlations among different features. This would aid in determining significant features we should be prioritizing. Since correlation values were small we had to lower our threshold value for this table. We chose anything >-.30 and .30<, since it ranges from -1 to 1. This is also important so we can exclude highly correlated features as it may lead to overfitting amongst features.

Education num and Income- 0.34
Hypothesis: As education level increases, income increases.
Marital status and Income- 0.39
Hypothesis: Married individuals will have a higher income.
Relationship and Income- 0:36
Hypothesis: Husbands will have a higher income than wives who will have a higher income than unmarried.



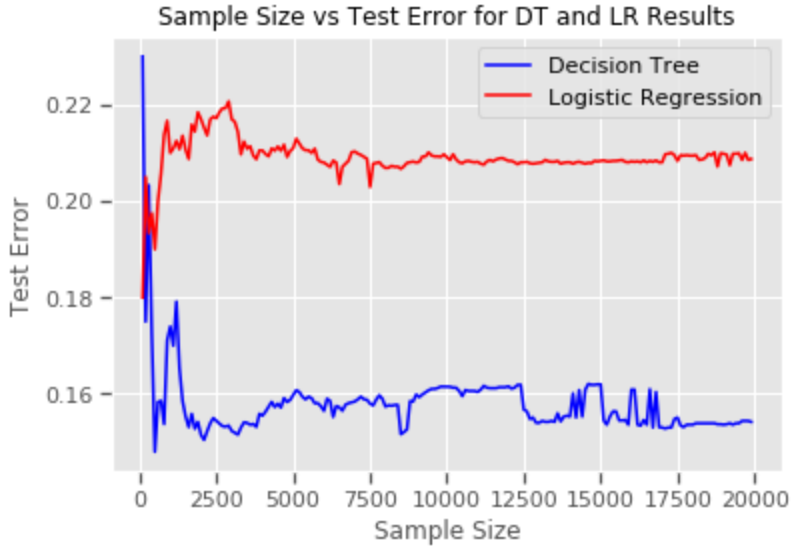
Heat map to identify correlations among different attributes in our dataset

Models and Algorithms

We chose to use Decision Trees and Logistic Regression as they are ideal for supervised machine learning. Decision Trees are great in this scenario since there are many categorical values we are accounting for and it only has 14 initial features. After encoding the labels the numbers of features range from 32 to 56. The number of features may play a role into determining the success of the algorithm and they are easy to overfit but we limit this by keeping

the depth at 5. Logistic Regression is also prime in this case since we need to classify a binary feature.

We split the dataset into test and train sets, with 80%(training) and 20%(testing) split.



Results and Analysis

The best way to determine the “fairness” of our model/dataset we first have to find the best performing model. In order to determine this we had few measures; accuracy, precision, specificity, recall, f1-score, overall ROC curve, and ROC-curve on significant features.
We had three data frames we performed our testing on, the Base, One-Hot-Encoded(OHE) and Binary-Encoded. Each of these data frames were tested with two different algorithms, Logistic Regression and Decision Trees.

Accuracy = (tp + tn) / (tp + tn + fp + fn)
Precision = (tp) / (tp+fp)
Recall = (tp) / (tp+fn)
Specificity = (tn) / (tn+fn)
F1-Score = (2 * (precision * recall)) / (precision + recall)

	datasetType	modelType	accuracy	precision	recall	specificity	f1-score
0	Base Model	Logistic Regression Model	0.784	0.723	0.273	0.790	0.396
1	Base Model	Decision Tree Model	0.845	0.807	0.531	0.853	0.641
2	One Hot Encoded Model	Logistic Regression Model	0.851	0.739	0.602	0.879	0.664
3	One Hot Encoded Model	Decision Tree Model	0.851	0.779	0.544	0.866	0.641
4	Binary Encoded Model	Logistic Regression Model	0.843	0.724	0.558	0.870	0.630
5	Binary Encoded Model	Decision Tree Model	0.840	0.731	0.530	0.863	0.614

Based on the accuracy of the models we observe the One Hot Encoded model with Logistic Regression performed the best out of all the models. However, we know that accuracy shouldn't be the only model when it comes to determining the performance of a model. We also noticed the Base Model with Decision Trees had the best precision score even though it was near the accuracy of the other models. The lowest performing model was the Base Model with Logistic Regression and that makes sense because it didn't have any feature engineering towards it. The Decision Tree seemed to perform the best on all of them because our dataset naturally had less features to analyze so it seemed the most plausible in this case. Additionally, the cardinality of the individual features never exceeded 15.

We know that recall gives us our true positive rate. In terms of our dataset, the recall will give us the number of people making more than $50,000 that our model predicts correctly over all the people making more than $50,000. Looking at this number from the scope of an individual as this number is something they would want to be maximized so it can and the model that does this best is the One-Hot-Encoded Model with logistic regression as it had the highest recall. It should be noted that recall in this case is the percentage of people that are actually making more than $50,000 how many of those are predicted to make more than $50,000.

Precision in this case would be the percentage of people that make more than $50,000 over the number of people our model predicted would make more than $50,000. This number should be high from the perspective of a business and the model that does this the best is the base model with the decision tree model.

The measures we have discussed so far seem incomplete and can tend to be favored depending on the point of view. The most balanced of all measures seems to be the f1-score as it is the harmonic mean between precision and recall. It emphasizes a balance between the two even if there is not one. The highest f1-score is the OHE model with Logistic Regression.

Going forward to assess the fairness of our model we will be performing tests on the base model with decision trees and OHE model with logistic regression as these two models had the highest f1-scores overall.

We began our analysis with our own **implicit bias** based on our initial data analysis. That is, we held assumptions that our model would incorrectly predict the negative class for female individuals more often than male individuals. We also assumed that it would be biased for the race category in the sense where it incorrectly produced more false negatives for all races that weren't white. In order to find out if our initial statements were true, we designed a method, split_on_val_eq, that allows us to split our test data into four different sets. Given a column and a feature value, our method pushed rows that met the criteria in one set, and their corresponding labels in another set. All other data was pushed into the two leftover sets.

Utilizing this method, we isolated all of the instances in the dataset where the gender was male and the instances where the gender was female. The results, to our surprise, were not what we expected.

The total number of false negatives for female individuals proportional to the size of the subset was much less than that of males in our test set. Furthermore, the accuracy of our model was much higher for males, as shown in Figure 5.

With further research, we realized that we should not look at the number of false negatives proportional to each group's size. We want to analyze the **proportion of positives that yield the negative outcome in the test**. That is, our evaluation metric should be conditional on whether the subgroup first meets the criteria that we're looking for. To test this, we looked at the *false negative rate* of each gender group.

After calculating the false positive rates on the base model with decision trees on each group, we quickly witnessed that the model produced a FNR of ~46.7% for males, and a FNR of ~60% for females. We also performed the same test on our OHE model with logistic regression and noticed the FNR for males was ~38.9% and a FNR of about ~51% for females. It made sense the better performing model lowered the FNR rate for both respective subgroups but there appears to be an inherent bias towards male. That could be attributed to the higher number of males that were recorded and the time frame the data was recorded from.

Our findings on only a couple subgroups of the dataset propelled us to iterate through all columns and their possible values to see which attributes and their respective values were producing the highest false negative rate. In order to achieve this, we outputted attributes and their corresponding values for which our model produced a higher FNR than a preset threshold. Since our main goal was to find subgroups that had a worse FNR than females, we initially chose 75% as our threshold value as that was the percentage of people that made <=50k in our total dataset. We showcase our observations below:

```
OHE Features with Significant FNR Threshold Values
----------------------------------------------------
4
Base Features with Significant FNR Threshold Values
----------------------------------------------------
61
```

Our OHE model had significantly less features and the features that were common in both had to do with relationship status/marital status. These two values can easily overlap one another so for future testing it makes sense to combine them into one attribute to prevent further correlated values.
When lowering our threshold to 50% we observed the following:

```
OHE Features with Significant FNR Threshold Values
----------------------------------------------------------
28
Base Features with Significant FNR Threshold Values
----------------------------------------------------------
129
```

Our OHE model still had significantly less features and the features we observed that were common amongst both pertained to: gender, race, relationship/marital status, education, and occupation.
These features appear to have the most bias towards them. We were aware the reporting bias would lead to some discrepancy between the two different genders and races. However, we did not expect the other features to be impacted. We attribute some of these biases in relationship status to some more reporting bias. We noticed individuals who were married had a lower FNR rate compared to those who weren't. This could be attributed to individuals combining their incomes with their spouses resulting in an overall higher income. Education and Occupation didn't originally cross our minds but after observing the individual categories its apparent those who had higher education levels and had more demanding jobs had lower FNR compared to the others.

Here are a few of our findings that we found particularly interesting:

Our base model had a FNR of ~78.7% when only operating on entries where the individual's highest level of education was 12th grade. There were 3,248 total instances in the test set that met this criteria.

Our model had a FNR of ~91% when only operating on entries where the individual's age was 26 years old. There were 247 total instances in the test set that met this criteria.

Our model had a FNR of ~66% when only operating on entries where the individual's native country is Mexico. There were 220 total instances in the test set that met this criteria.

Overall, our base model had a higher rate of FNR for most features and using our OHE model helped us cross-verify which features were actually biased since our OHE model balanced precision and recall the best.

Conclusion/Future Work

Based on our analysis we learned making models is a constant cycle. We start by understanding our domain, then choosing the right initial algorithms and constantly reassessing our domain to achieve the best performing models. Given the time constraint of this project, we conducted 3 cycles of this process, thus resulting in our 3 separate dataframes. In the future, we would like to feature engineer relationship/marital status as both of these attributes classified as

significant values by our threshold testing. After this, we would be able to remove the high correlation between these features, identified in our correlation heatmap. Finally, we would like to add another algorithm such as Naive Bayes classifier since it is commonly used in classification problems.

Through our project we learned that the results of the model can be viewed differently depending on the use of the data predictions. If we want to favor the business/companies, then we would choose the model with the highest precision. It is tough to get both high precision and high recall as in most cases if precision is high then recall is low, but this would yield the greatest percentage companies/businesses desire. The recall metric would be ideal for the customer as it delivers high recall but lower precision and won't exclude any individuals. Higher recall would increase the chance they are classified to make more than 50k. If we want to assess the best model from the point of views of the people and the companies/businesses they interact with, then our OHE model with Logistic Regression would be the best since it is the most "fair" of everything we have so far.

Contribution

Rohan - Did research on dataset and income prediction applications. Researched applications of fairness as it pertains to the project (instances of bias in our dataset and how they pertain to common issues with fairness in ML). Worked on data preprocessing for the initial dataframe and analyzed the Logreg and DT models for accuracy with varying sample sizes. Analyzed the FNR for different models and compared results. Converted the Docs write up to LaTeX.

Kartik - Performed Data Preprocessing, Feature Engineering, and Data Analysis. Made OHE(One-Hot-Encoded) and BE(Binary-Encoded) dataframes. Made robust and scalable methods to account for various Models and their respective performance metrics to assess our various models and dataframes. Summarized Feature Engineering/Initial Data Analysis in docs. Analyzed FNR without various thresholds on all models.

Radhika - Researched the domain space and metrics for fairness in our project. Performed Data Preprocessing and Feature Engineering for BE(Binary-Encoded) model and made respective graphs. Summarized model performance metrics in Google Docs and converted respective portions to LaTeX. Created and edited final presentation with iMovie.

References

https://towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040cb

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

https://towardsdatascience.com/beyond-accuracy-precision-and-recall-3da06bea9f6c

Angwin, Julia, et al. “Machine Bias.” ProPublica, 9 Mar. 2019, www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

“UCI Machine Learning Repository: Adult Data Set.” UCI Machine Learning Repository: Adult Data Set, Center for Machine Learning and Intelligent Systems, 2020, archive.ics.uci.edu/ml/datasets/Adult.

Zhong, Ziyuan. “A Tutorial on Fairness in Machine Learning.” Medium, Towards Data Science, 27 July 2019, towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040cb.

Will Koehrsen. 2018. Beyond Accuracy: Precision and Recall. (March 2018). Retrieved March 15, 2020 from https://towardsdatascience.com/beyond-accuracy-precision-and-recall-3da06bea9f6c