Machine learning is producing increase advantages in business values and efficiencies. However, we need to be alert of bias in machine learning models that are producing inequalities to large portions of the population. These bias models can target gender, race, age, income levels, … The effect can include lost opportunities for employment, financial services, housing, fair judicial system, …

This bias and potential inequality can be an unnoticed process but have a powerful impact. It is up to us to include in our development and maintenance process to look for bias in eradicate it.

Machine learning by default is bias, since it relies on statistical bias. This is required to make predictions, classifications, and correlations on new data the model has never seen before. However, focus needs to be put on the bias on the algorithms and training data used to create the models in the first place.

We will focus on  [research conducted by ProPublica](https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing), a non-profit research institution, it was found that COMPAS, a machine learning algorithm used to determine criminal defendants’ likelihood to recommit crimes.

**Outline:**

1. Get data
2. Initial - Exploratory data analysis (EDA)
3. Initial – Data Wrangling
4. Visual Exploratory data analysis (EDA) – that highlights some bias
5. Feature Engineering - Prepare the data for Machine Learning Algorithms
6. ReCreate a Compass Regressor model
7. Apply Variant Fair Autoencoder (using tensorflow) to minimize bias
8. Conclusion

**Data:**

* Compass dataset - The data set tracks Broward county Florida
* US census data for some initial data comparisons

**[Code link](https://github.com/rivasjmr/Springboard/blob/master/COMPASS/Bia_Compass_dataset_using_VFAE_v2.ipynb)**

**EDA – Data Bias**

Broward county Florida [demographics](https://www.census.gov/quickfacts/fact/table/browardcountyflorida/RHI125216#viewtop)

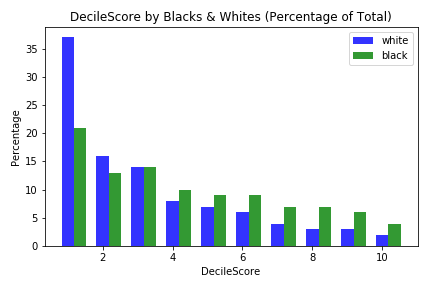
|  |  |  |
| --- | --- | --- |
| **Race** | **% demographics distribution** | **% data distribution** |
| White | 37.7 | 35.8 |
| African American | 29.7 | 44.5 |
| Hispanic or Latino | 28.7 | 14.36 |
| Asian | 3.8 | 0.53 |
| Other | 0.1 | 4.81 |

|  |  |  |
| --- | --- | --- |
| **Gender** | **% demographics distribution** | **% data distribution** |
| Male | 48.7 | 78.09 |
| Female | 51.3 | 21.91 |

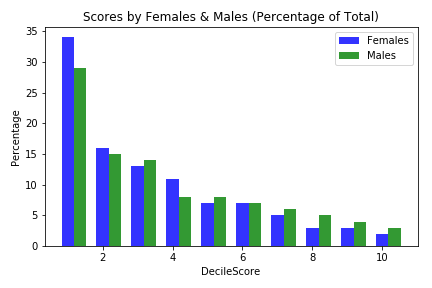
*The data distribution does not reflect demographics distribution*

**DecileScore is between 1 thru 10. The higher the score indicates higher risk.**

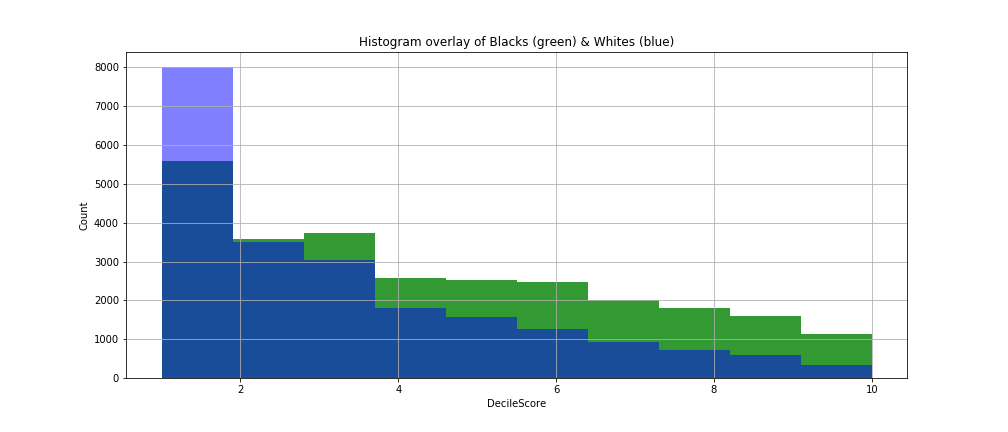
|  |  |
| --- | --- |
| **Race** | **Mean DecileScore** |
| African American | 4.29 |
| White | 3.11 |
| Hispanic or Latino | 2.86 |
| Asian | 2.37 |
| Other | 3.1 |

**

|  |  |
| --- | --- |
| **Gender** | **Mean DecileScore** |
| Male - 3.66 | 3.66 |
| Female - 3.26 | 3.2 |

**

*The DecileScore (risk) are higher for African American than all other races and higher for males than females*

**

**Data Wrangling**

Clean up date set

Create groups

Sub data sets

**Feature Engineering**

Created new column Age from date of birth

Use Label Encoder and One-hot Encoder

**Recreate Compass Model**

Used MLPRegressor

**Apply Variant Fair Autoencoder (using tensorflow) to minimize bias**

**Conclusion**

We might lose some accuracy in minimizing bias, it is important to be aware of the bias and have methods to minimize it