Henil Vedant MS CYBERSECURITY 495670888

Using Microsoft Fair-learn to make Fair Decisions.

Project Report

CIS700: Machine Learning for Privacy and Security

Syracuse University

Cyber Security

Henil Vedant

hhvedant@syr.edu

Henil Vedant MS CYBERSECURITY 495670888

1. Abstract

Negative attitudes (e.g. stereotypes, prejudice) and unintentional biases (e.g. organisational practices or internalized stereotypes) can provoke a discriminative behavior in humans. AI systems trained on datasets that include biased human decisions will learn making biased decisions rather than fair decisions. When dealing with sensitive features (e.g. gender, age, nationality, etc.), it is crucial to to ensure fairness of a model before deploying it in a real-world environment.

Our success, happiness, and wellbeing are never fully of our own making. Others' decisions can profoundly affect the course of our lives: whether to admit us to a particular school, offer us a job, or grant us a mortgage. Arbitrary, inconsistent, or faulty decision-making thus raises serious concerns because it risks limiting our ability to achieve the goals that we have set for ourselves and access the opportunities for which we are qualified.

Microsoft open-source Python package called fairlearn can be used for analyzing the model's fairness and mitigating any observed unfairness issues.

2. Brief Introduction to FAIRNESS in Fair-Learn:

Fairlearn, an open source toolkit that empowers data scientists and developers to assess and improve the fairness of their AI systems. Fairlearn has two components: an interactive visualization dashboard and unfairness mitigation algorithms. These components are designed to help with navigating trade-offs between fairness and model performance. Prioritizing fairness in AI systems is a sociotechnical challenge. Because there are many complex sources of unfairness— some societal and some technical—it is not possible to fully "debias" a system or to guarantee fairness; the goal is to mitigate fairness-related harms as much as possible.

Model bias detection: Supported fairness metrics:

Demographic Parity: This metric states that the proportion of each segment of a protected feature (e.g. gender) should receive the positive outcome at equal rates.

Equalised Odds: This metric states that the model should correctly identify the positive outcome at equal rates across groups, but also miss-classify the positive outcome at equal rates across groups (creating the same proportion of False Positives across groups).

***Model bias mitigation** :* Model estimation approaches:

Exponentiated Gradient Reduction: The reduction approach for achieving fairness in a binary classification setting. Underlying classification method is treated as a black box.

Grid Search: The reduction approach for achieving fairness in a binary classification and regression settings. If the protected attribute is non-binary, then grid search is not feasible.

Post-processing approaches: *Threshold Optimizer:*

The postprocessing approach in which the classifier is obtained by applying group-specific thresholds to the provided estimator. The thresholds are chosen to optimize the provided performance objective (e.g. accuracy score) subject to the provided fairness constraints (e.g. demographic parity).

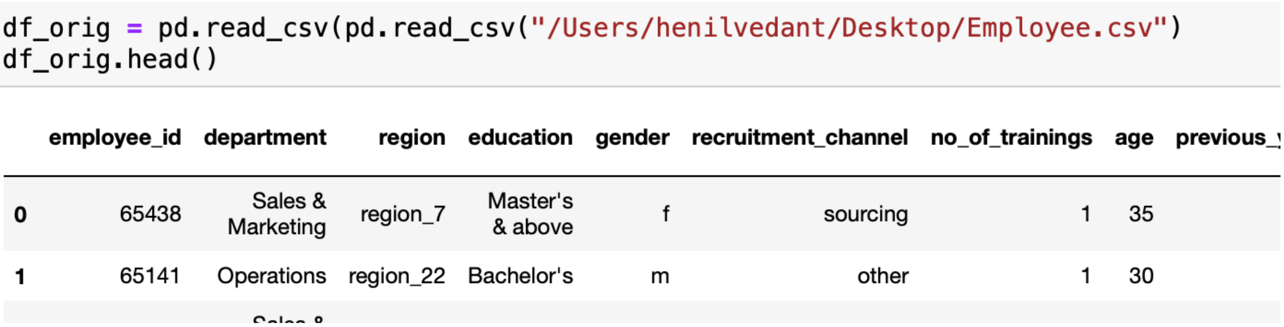
3. Proposed Work

I will use the Employee HR Promotion dataset to predict employees that are likely to be promoted based on their personal and performance parameters. This is the binary classification task with the target `is promoted`. The target can be equal to 0 or 1. When `is promoted` is equal to 0, an employee is not promoted. Otherwise, an employee is considered for the promotion. The dataset contains potentially sensitive features, such as gender, age, region and department. Comparing the results, we get using the fairness unaware and the fairness aware ML models we build observation and Results.

4. Experiment & Results

Data pre-processing:

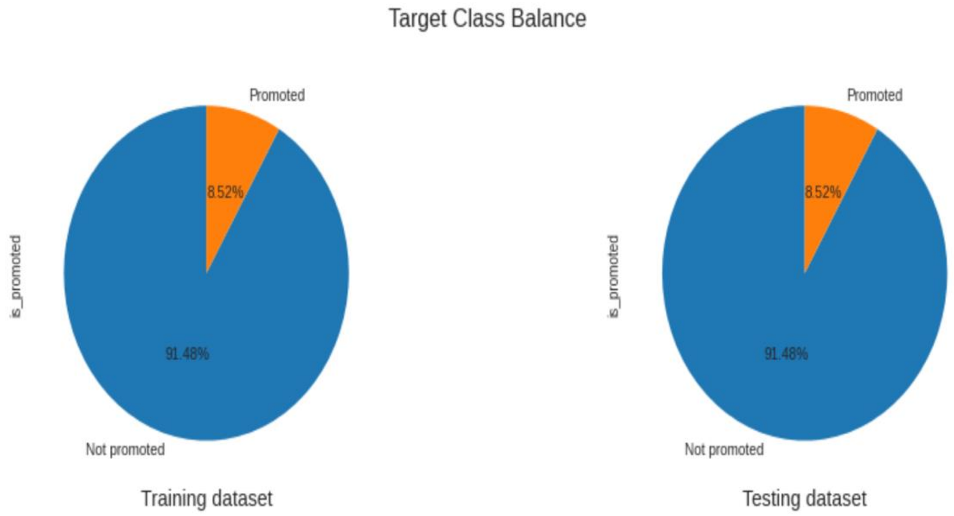
HR dataset has employee details as listed: ID - Unique identifier of each employee, Dept name - states which department the employee belongs too (HR, Finance, Tech), location - this shows the base location of each employee), Travel required - this tells if the employees job requires him/her to travel to other places (Yes/No). We encode categorical features: department, region, education, gender, recruitment channel and impute missing values.



```
df_orig = pd.read_csv(pd.read_csv("/Users/henilvedant/Desktop/Employee.csv")
df_orig.head()
```

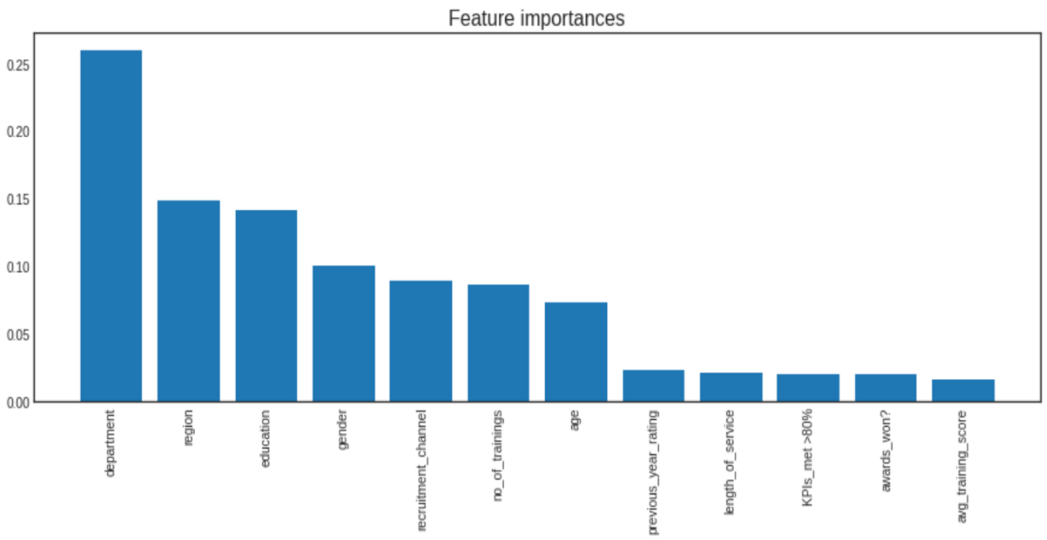
	employee_id	department	region	education	gender	recruitment_channel	no_of_trainings	age	previous_
0	65438	Sales & Marketing	region_7	Master's & above	f	sourcing	1	35	
1	65141	Operations	region_22	Bachelor's	m	other	1	30	

To split the promotion dataset into training and testing sets, we will use train test split function with stratify=y. The stratify parameter makes a split so that the proportion of values in the training and testing sets is the same as the proportion of values in y (i.e is promoted). As above, y (is promoted) is a binary categorical variable and there are 91.48% of zeros and 8.52% of ones. Thus, stratify=y will make sure that a random split has 91.48% of 0's and 8.52% of 1's.



Prediction Model:

Let's start with building a fairness unaware ML model. The imbalanced-learn library provides a classifier called Balanced Bagging Classifier that implements a random under sampling strategy on the majority class within a bootstrap sample in order to balance the two classes. The global importance of features used by the model. To do so, we will calculate the mean importance across the estimators of the Balanced Bagging Classifier.



The fairlearn package provides fairness-related metrics that can be compared between groups and for the overall population. The goal is to assure that neither of the departments has substantially larger false-positive rates or false-negative rates than the other groups.

By doing that we enforce parity not just on selection rates but also on the rejection rates that we are going to be enforcing on this protected attribute department.

Therefore, as a protected (sensitive) feature we will set department that has a highest impact on predictions of the trained model.

Using existing metric definitions from scikit-learn we can evaluate metrics to get a group summary. As the overall performance metric we will apply the area under ROC curve (AUC), which is suited to classification problems with a large imbalance between positive and negative examples. As the fairness metric we will use equalized odds and demographic parity.

Results:

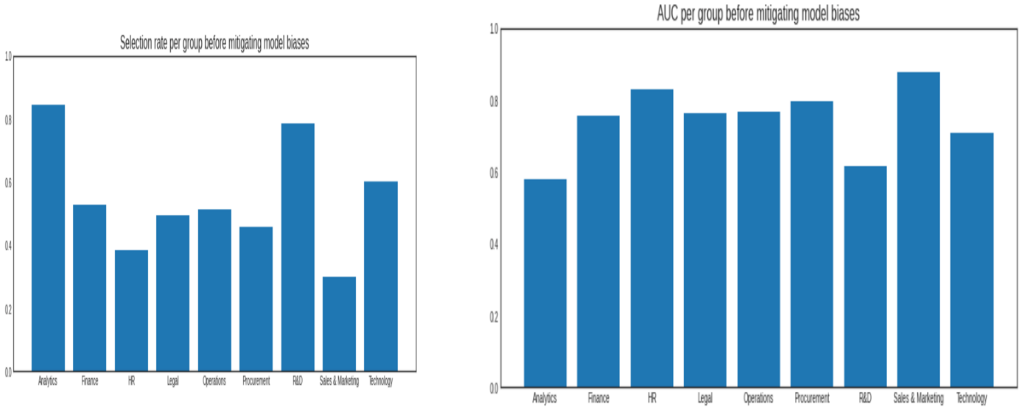


Fig1 (a) – left Fig1(b) - right

Fig1(a) and (b) are the results for an unfair mode.

It can be noticed that some of the departments have significantly higher AUC values and higher selection rate in comparison to others.

We can conclude that the model is likely to overpredict (predict 1 when the true label is 0) and underpredict (predict 0 when the true label is 1). Therefore, the application of such model in a real-world environment may cause discrimination of employees based on departments. And in order to avoid and remove this discrimination we can use existing metrics and bias mitigation techniques from Microsoft fairlearn and create a new model from the scratch which will take care of the selection rates.

Achieving non-discrimination with Fairlearn :

The fairlearn package supports different techniques for mitigating model bias.In our case, we will create a new model while specifying the fairness constraint called Demographic Parity, because there are disparities between groups of department. The Exponentiated Gradient mitigation technique developed by Microsoft fits the provided classifier using Demographic Parity as the objective, leading to a vastly reduced difference in selection rate.
Since Balanced Bagging Classifier of imbalanced-learn is not yet supported in fair learn, we will use LGBM Classifier as a base model.

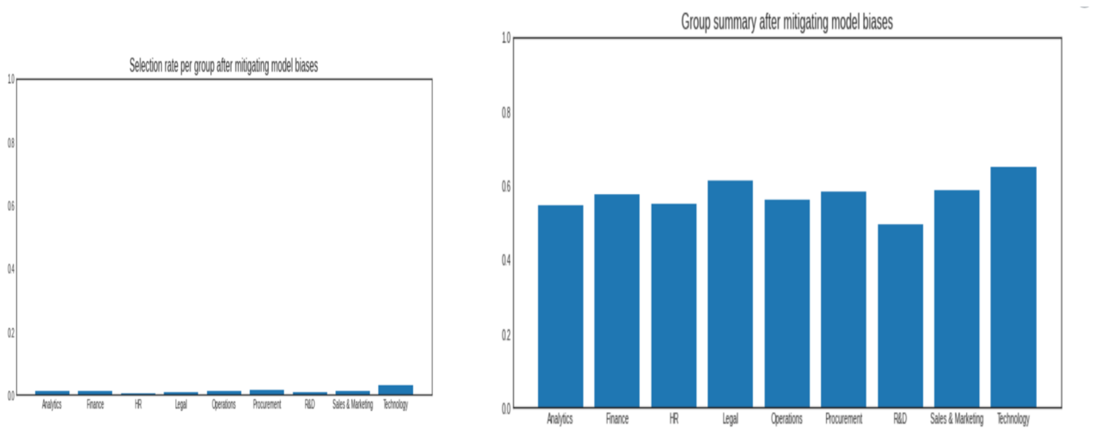


Fig2 (a) – left Fig2(b) – right Fig2. (a) and (b) are the results for a fair mode.

The Exponentiated Gradient algorithm significantly reduces the disparity according to multiple metrics. However, the performance metrics (balanced error rate as well as AUC) get worse. Before deploying such a model in practice, it would be important to analyze why we observe such a performance decrease. One of the reasons might be the lack of informativeness of available features for one of the groups of a protected feature department.
Note that unlike the unmitigated model, Exponentiated Gradient produces 0/1 predictions, so its balanced error rate difference is equal to the AUC difference, and its overall balanced error rate is equal to 1 - overall AUC.

5. Observation, Conclusion and Future work:

We learn that fairness is task specific, and it is important to consider ethical and social responsibilities while designing a fair and responsible system. The objective of the algorithm

and the subjective constrain of fairness introduce a trade-off in system between the accuracy and utility (performance) of the system. The constraints of fairness need to be crafted by responsible people from social and technological domain. The impact of imposing the above constraints on the accuracy truly depends on the dataset, the fairness definition used as well as the algorithms used. In general, however, fairness hurts accuracy because it diverts the objective from accuracy only to both accuracy and fairness.
Therefore, in reality, a trade-off should be made. There are many algorithms that claim to help improve fairness. Most of them fall into three categories: preprocessing, optimization at training time, and post-processing.

In practice, without surprise, the method that achieves the best trade-off between accuracy and fairness is via optimization at training time. However, preprocessing and post-processing methods grant the ability to preserve fairness without modifying the classifiers. Such feature is desirable when we do not have power to modify the classifiers.
Observational criteria can help discover discrimination but are insufficient on their own. Causal viewpoint can help articulate problems, organize assumptions.
Any two of the three group fairness definitions demographic parity, equalized odds, and predictive rate parity cannot be achieved at the same time except in degenerate cases.
Trade-off between accuracy and fairness usually exists. There are three streams of methods: preprocessing, optimization at training time, and post-processing. Each has pros and cons.
Most fair algorithms use the sensitive attributes to achieve certain fairness notions. However, such information may not be available.

Are There Any Real-World Consequences for Not Developing Fairness-Aware Practices?

It is also interesting to answer this question and understand the consequences of deployed unfair systems and how the algorithmic decision-making process affects the promotion of employees in their organization due to discrimination of the alogirthm and not due to the undue merit of an individual, such a setting generally extends the idea of how fair are the systems that are deployed, how accurate are the decisions based on these systems, is there a way to remove/deal with societal injustice, the members of a particular distribution may have faced prior discrimination in real life and in participating datasets subsequently. The practices which were undertaken before need to be assessed and taken into consideration to understand the true fairness implication.

6. References

Fairness Through Awareness- Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer

Reingold, Rich Zemel.

[AAL] AALIM. http://www.almaden.ibm.com/cs/projects/aalim/.

[AAN+98] Miklos Ajtai, James Aspnes, Moni Naor, Yuval Rabani, Leonard J. Schulman, and Orli Waarts. Fairness in scheduling. Journal of Algorithms, 29(2):306–357, November 1998.

[ABC+05] IttaiAbraham,YairBartal,HubertT.-H.Chan,Kedar Dhamdhere,Anupam Gupta, JonM. Kleinberg, Ofer Neiman, and Aleksandrs Slivkins. Metric embeddings with relaxed guarantees. In FOCS, pages 83–100. IEEE, 2005.

[BS06] Nikhil Bansal and Maxim Sviridenko. The santa claus problem. In Proc. 38th STOC, pages 31–40. ACM, 2006.

[Cal05] Catarina Calsamiglia. Decentralizing equality of opportunity and issues concerning the equality of educational opportunity, 2005. Doctoral Dissertation, Yale University.

[CG08] T.-H. Hubert Chan and Anupam Gupta. Approximating TSP on metrics with bounded global growth. In Proc. 19th Symposium on Discrete Algorithms (SODA), pages 690–699. ACM-SIAM, 2008.

[CKNZ04] Chandra Chekuri, Sanjeev Khanna, Joseph Naor, and Leonid Zosin. A linear programming formulation and approximation algorithms for the metric labeling problem. SIAM J. Discrete Math., 18(3):608–625, 2004.

[DMNS06] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In Proc. 3rd TCC, pages 265–284. Springer, 2006

Fairness through awareness slides – Moritz Hardt.

Dataset and libraries – Microsoft fairlearn.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirch- ner. Machine bias. ProPublica, May, 23:2016, 2016.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness and Machine Learning. fairmlbook.org, 2019. http://www.fairmlbook.org.

Flavio P Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varsh- ney. Optimized pre-processing for discrimination preven- tion. In Proceedings of the 31st International Conference on Neural Information Processing Systems, pages 3995–4004, 2017