Using SageMaker to deploy a Fair Binary Predictor

Project's Domain Background

The project's domain background is about binary classification and also touches on AI fairness [1].

Database and Inputs

COMPAS, short for *Correctional Offender Management Profiling for Alternative Sanctions*, is a case management and decision support tool developed and owned by Northpointe (now Equivant) used by US courts to assess the likelihood of a defendant become a repeat offender (recidivism). COMPAS in the USA has already assessed the risk of more than 1 million defendants since its development in 1998. In May 2016, writing for the investigative journal ProPublica [2], analyzed the effectiveness of COMPAS on more than 7,000 prisoners in the Broward County, Florida, between 2013 and 2014. The data can be found in the following GitHub repository: https://github.com/propublica/compas-analysis. And the methodology employed by ProPublica [3] is also available. First, I will filter out rows where `days_b_screening_arrest` is over 30 or under -30, leaving the data with 6,172 instances. Then, following [4], I will also subset the observations only for blacks and whites instances. This leaves the data with 5,278 instances.

Solution Statement

I will be using SageMaker's XGBoost algorithm with the intent of maximizing precision score (MAP – Mean Average Precision). This is necessary to tackle the unbalanced false positive rate between black and white offenders explained in the next section.

Benchmark Model

ProPublica analysis indicated that the predictions were unreliable and were racially biased. The overall accuracy of COMPAS for white defendants is 67%, just slightly higher than the accuracy of 63.8% for black defendants. The mistakes made by COMPAS, however, affected black and white defendants differently: black defendants who did not recidivist were incorrectly predicted to recidivist (false positives) at a rate of 44.9%, almost double the number of whites in 23.5%. In other words, COMPAS scores appeared to favor white defendants over black defendants, underestimating the recurrence of whites and repeat offenders of black defendants. Because of this issue of racial discrimination, COMPAS was used by many machine learning researchers to propose and test techniques for how to make fairer algorithms [4]–[13]. COMPAS has become the "gold standard" for validating the performance of techniques that counteract machine learning algorithms.

An outline of the project design

The model will be trained by deploying a notebook instance in SageMaker and calling XGBoost training jobs from that instance. I will also call hyperparameter tuning jobs to maximize MAP (Mean Average Precision) to try to overcome COMPAS original low precision for black offenders. The model will be tested by batch transform jobs. All of these jobs will be called from the notebook instance's Jupyter Notebook. Finally, the whole project will be communicated and described in a blog post on Medium.

References

[1] S. Barocas, E. Bradley, V. Honavar, and F. Provost, "Big Data, Data Science, and

Civil Rights," 2017. Accessed: Sep. 22, 2019. [Online]. Available: http://arxiv.org/abs/1706.03102.

[2] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine Bias: there's software used across the country to predict future criminals. And it's biased against blacks," *ProPublica*, 2016.

[3] J. Larson, S. Mattu, L. Kirchner, and J. Angwin, "How we analyzed the COMPAS recidivism algorithm," *ProPublica*, 2016.

[4] A. Chouldechova, "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments," *Big Data*, vol. 5, no. 2, pp. 153–163, Jun. 2017, doi: 10.1089/big.2016.0047.

[5] J. Dressel and H. Farid, "The accuracy, fairness, and limits of predicting recidivism," *Sci. Adv.*, vol. 4, no. 1, p. eaao5580, Jan. 2018, doi: 10.1126/sciadv.aao5580.

[6] N. Grgic-Hlaca, M. B. Zafar, K. P. Gummadi, and A. Weller, "The Case for Process Fairness in Learning: Feature Selection for Fair Decision Making," in *Symposium on Machine Learning and the Law at the 29th Conference on Neural Information Processing Systems*, 2016.

[7] M. B. Zafar, I. Valera, M. G. Rodriguez, and K. P. Gummadi, "Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment," in *26th International World Wide Web Conference, WWW 2017*, 2017, pp. 1171–1180, doi: 10.1145/3038912.3052660.

[8] A. K. Menon and R. C. Williamson, "The cost of fairness in binary classification," in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 2018, vol. 81, pp. 107–118, [Online]. Available: http://proceedings.mlr.press/v81/menon18a.html.

[9] E. Pierson, S. Corbett-Davies, and S. Goel, "Fast Threshold Tests for Detecting Discrimination," Feb. 2017, Accessed: Sep. 22, 2019. [Online]. Available: http://arxiv.org/abs/1702.08536.

[10] S. Corbett-Davies and S. Goel, "The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning," Jul. 2018, Accessed: Sep. 22, 2019. [Online]. Available: http://arxiv.org/abs/1808.00023.

[11] A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach, "A Reductions Approach to Fair Classification," 2018.

[12] R. Berk *et al.*, "A Convex Framework for Fair Regression," Jun. 2017, Accessed: Sep. 22, 2019. [Online]. Available: http://arxiv.org/abs/1706.02409.

[13] J. Fitzsimons, A. Al Ali, M. Osborne, and S. Roberts, "A General Framework for Fair Regression," *Entropy*, vol. 21, no. 8, p. 741, Jul. 2019, doi: 10.3390/e21080741.