The metrics Precision and Recall are often used, especially in text classification and information retrieval. **Recall is the same as true positive rate,** while **precision is TP/(TP+ FP),** which is the accuracy over the cases predicted to be positive. The **F-measure is the harmonic mean of precision and recall at a given point,** and is:

$$F-measur = 2\cdot\frac{precision\cdot recall}{precision + recall}$$

Practitioners in many fields such as statistics, pattern recognition, and epidemiology speak of the sensitivity and specificity of a classifier:

Specificity = $TP/(TP+FN) =$ **True positive rate**

Sensificity = $TN/(TN+EP) =$ **True negative rate = 1-False positive rate**

You may also hear about the positive predictive value, which is the same as precision. Accuracy, as mentioned before, is simply the count of correct decisions divided by the total number of decisions, or:

$$Accuracy = \frac{TP+TN}{P+N}$$

6.1.3 Visualizing Model Performance (Ch. 8) For example:

A Describe a ranking classifier.

This introduces a complication for which we need to extend our analytical framework for assessing and comparing models. **"The Confusion Matrix" on page 189 stated that a classifier produces a confusion matrix.** With a **ranking classifier, a classifier plus a threshold produces a single confusion matrix.** Whenever the threshold changes, the confusion matrix may change as well because the numbers of true positives and false positives change.

Figure 8-1 illustrates this basic idea. As the threshold is lowered, instances move

Reading 8.2 Institute of International Finance (May 2019). Machine Learning Thematic Series Part II: Bias and Ethical Implications.

8.2.1 Bias and Ethics in Machine Learning For example:

A Discuss and contrast the statistical and social definitions of bias.

What is bias? Bias has contradictory meanings, as the history of the word itself indicates, and can have various interpretations in different contexts. The statistical definition of bias relates to systematic differences between a population subsample and the whole population at large, where selection bias occurs when the set of data inputs to a model is not representative of a population, and results in conclusions that could favor certain groups over others. In ML systems models are trained to predict the future based on the past, as such what the ML model learns depends on the data used to train it.

In contrast, the popular, or social, definition of bias is human judgement made based on preconceived notions or prejudices, as opposed as the impartial evaluation of facts. Society tries to address this with legislation to prevent discrimination against particular groups of people. Protected groups vary by jurisdiction.

Both the statistical and the social meanings of bias are relevant and may lead to discrimination. Laws typically evaluate the discrimination using two distinct notions: disparate treatment, and disparate impact.

B Explain the trade-off between variance and bias in models.

Variance refers to the amount by which the model prediction would change if it were to be estimated with different training data. Bias refers to the error

8.2.2 Types of Bias in Machine Learning For example:

A Explain how machine learning methods can increase or decrease discrimination and prejudices in models.

What has changed is that ML has created the potential for machines to learn from data that reflects human biases, including unconscious ones, and then exhibit and perhaps even amplify those biases. The main concern, and the mo tivation for the increased scrutiny of bias in the world of ML, is that misguide d correlations could have powerful implications given the automated nature of ML algorithms, and inherent biases can produce data that amplifies the biases already present in society. Additional wider aspects related to social "perception" of ML algorithms might also come into play.

Financial institutions are taking a cautious approach to the use of ML. Data protection, security and integrity are already a key part of the design process for banks. Statistical models used for credit decisioning are already subject to model governance, model risk frameworks, and fair lending assessments. Financial institutions can adapt current governance and risk management frameworks to develop approaches to ensuring the ethical use of new technologies such as ML.

For instance, during a fair lending assessment in the U.S. examiners obtain a list of the variables considered and may conduct a comparative analysis between approved and denied applications to examine whether there are indications of disparate treatment. Many firms use existing regulations and supervisory guidance as a starting point, and carefully determine how to adapt these processes to the use of new technologies.

One of the biggest concerns firms share are with biased data leading to biased algorithms, in particular related to protected/sensitive features that could create moral, ethical and legal problems.

Reading 8.3 Das S., M. Donini, J. Gelman, K. Haas, M. Hardt, J. Katzman, K. Kenthapadi, P. Larroy, P. Yilmaz, and M. B. Zafar (2021). Fairness Measures for Machine Learning in Finance. The Journal of Financial Data Science, 3(4): 33-64. Only pages 33-50 from this reading will be used for the FDP exam.

8.3.1 Algorithmic Biases and Finance For example:

A List the three broad approaches to fairness-aware machine learning (FAML).

There are three broad approaches to FAML: (i) Methods that try to manage biases in the data used for training, (ii) methods that impose fairness during training, and (iii) methods that mitigate bias post-training.

B Explain the practical challenges in FAML, including where bias appears in models and how metrics of fairness may conflict with other metrics.

Practical Challenges in FAML: The trouble with implementing FAML is that there are too many notions of fairness and a lack of clarity on the prioritization of these definitions. We hope to catalog most of these measures for the finance domain, connect them to industry terminology and regulation, and resolve some conflicting definitions, but not all. At first glance, it would be comforting to assume that the more measures of fairness we use, the better, so that all types of fairness are imposed in, say, a lending algorithm. However, as we make an algorithm fair on one measure, it may become unfair on another, since the commonly used metrics for fairness often conflict with each other. Berk et al. (2017) present a set of six comprehensive measures of fairness. This paper shows that it is mathematically difficult to mitigate bias all measures. The intuition is that since there are just four numbers in the confusion matrix, and many more metrics, it is very hard to change these four numbers in really a fact and not a joke, satire, or comedy. More importantly, carefully consider expert credentials, and carefully check for confirmation bias.

6. Unanticipated machine decisions. Untrammeled machine-learning often arrives at optimal solutions that lack context, which cannot be injected into the model objective or constraints. For example, a ML model that takes in vast amounts of macroeconomic data and aims to minimize deficits may well come up with unintended solutions like super-normal tariffs leading to trade wars. This inadmissible solution arises because the solution is not excluded in any of the model constraints. The model generates untenable answers because it does not have context .

B Recognize and explain class imbalance and conditional demographic disparity in labels (CDDL).

Class imbalance (CI): Bias is often generated from an under-representation of the disadvantaged group in the dataset, especially if the desired "golden truth" is equality across groups. As an example, algorithms for granting small business loans may be biased against women because the historical record of loan approvals contains very few women, because women did not usually apply for loans to start small businesses. This imbalance can carry over into model predictions.

We will report all measures in differences and normalized differences. Since the measures are often probabilities or proportions, we want the differences to lie in

(−1, +1). We define **CI** $= \frac{n_a - n_d}{n}$ **$\in (-1, +1)$ in normalized form.** We see that **CI can also be negative, denoting reverse bias.**

Mostly, the proportion difference is what is needed, but sometimes we may need Mostly, the proportion difference is what is needed, but sometimes we may need the normalized difference, as in the case of the 80% Rule, that may be used to measure certain types of employment discrimination, see the 80% rule. In this case, it is the ratio that is important, so the normalized probabilities are able to capture this.

Conditional Demographic Disparity in Labels (CDDL): Wachter et al. (2020) developed this measure, which can be applied pre-training and also post-training.

negatives for each class.

$$\text{DRR} = \frac{TN_d}{\hat{n}_d^{(0)}} - \frac{TN_a}{\hat{n}_a^{(0)}}.$$

Also, just as DCA is related to DUA, we see that DRR is related to DCR .

8.3.3 Bias Mitigation For example:

A List and explain four methods of bias correction and mitigation.

Bias corrections can take many forms and may lead to different tradeoffs between fairness and accuracy for each ML model. Some common corrections that may be applied are as follows:

1. Removal of the class variable form the feature set. For example, restricted characteristics such as gender, race/ethnicity, and age may be part of the feature set and removal of these will mitigate some or all of the bias metrics mentioned above. However, as is to be expected, this will also impact accuracy. Moreover, the real problem often lies elsewhere. given that protected attributes arc almost always eliminated from feature sets, but not all features that are correlated with the attribute.

2. Rebalance the training sample pre-training. This corrects unfairness from differences in base rates. Synthetically increase the number of observations

$$n_d^{(1)} \text{ if } n_a^{(1)} \rangle n_d^{(1)}.$$

Synthetic oversampling is undertaken using standard algorithms such as SMOTE, available in SkLearn. Likewise, decrease $n_d^{(0)}$ if $n_a^{(0)} \langle n_d^{(0)}$. Both these corrections are in the spirit of affirmative action. Both these adjustments are intended to result in equal sized classes across (0, 1) labels. **If the class variable truly matters then rebalancing usually results in a loss in accuracy.** Random perturbation of class labels is also possible instead **of using oversampling.** But, this approach results in **different results every time.** One can also transform the features such that their joint distribution without the class variables remains more or less the same. but the correlation with the class variable is reduced. as close to zero as possible.

3. **Adjust labels on the training dataset and retrain.** For the advantaged class, **adjust the ground truth** such that $y_a=0$ if $\hat{y}_a=1$ and $p_a(X) < H + ŋ$, for some well define hyperparameter ŋ. That is, downgrade some of the borderline positive labels for the advantaged class. Likewset $y_a = 1$ if $\hat{y}_a = 0$ and $p_d(X) > H - ŋ$. i.e., upgrade some of the borderline negative labels for the disadvantaged class. Then re-run the ML, model fit. Recompute the various bias and accuracy metrics.

4. Adjust cutoffs post-modeling. The cutoff probability is usually set at H = 1/2. If bias is present, then the cutoff for the advantaged class can be adjusted to H + δ, and the cutoff for the disadvantaged class will be reduced to H - δ. This will change the predicted counts $\hat{n}_a^{(0)}$, and $\hat{n}_a^{(1)}$, $n_d^{(0)}$, $n_d^{(1)}$, change many of the bias measures as well as the accuracy of the model. Hyperparameter δ can be tuned appropriately, until a desired level of fairness and/or accuracy is achieved. The legal milieu may not accommodate direct alteration of the predictions, so the availability of this mitigation is subject to the domain of application.

These bias corrections will result in changes in fairness and accuracy for all of the ML models that are applied to the and there is a tension amongst them, we train our models with these fairness metrics as constraints, either applied ex-post or at training time. With multiple constraints, we need to either (ⅰ) choose one constraint (which is limiting), (ⅱ) weight the constraints to consolidate them into a single constraint, or (ⅲ) apply a min-max criterion, i.e., minimize the maximum bias metric under all the different constraints we choose to include while training the model.